

# Northumbria Research Link

Citation: Linden, Audrey H (2019) Heterogeneity of research results: New perspectives on psychological science. Doctoral thesis, Northumbria University.

This version was downloaded from Northumbria Research Link:  
<http://nrl.northumbria.ac.uk/id/eprint/42041/>

Northumbria University has developed Northumbria Research Link (NRL) to enable users to access the University's research output. Copyright © and moral rights for items on NRL are retained by the individual author(s) and/or other copyright owners. Single copies of full items can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided the authors, title and full bibliographic details are given, as well as a hyperlink and/or URL to the original metadata page. The content must not be changed in any way. Full items must not be sold commercially in any format or medium without formal permission of the copyright holder. The full policy is available online: <http://nrl.northumbria.ac.uk/policies.html>



**Northumbria  
University**  
NEWCASTLE



**UniversityLibrary**

---

# Heterogeneity of research results: New perspectives on psychological science

---

Audrey H. Linden

PhD

2019

---

Heterogeneity of research results: New  
perspectives on psychological science

---

Audrey H. Linden

A thesis submitted in partial fulfilment of  
the requirements of the University of  
Northumbria at Newcastle for the degree  
of Doctor of Philosophy

Research undertaken in the Faculty of  
Health and Life Sciences

September 2019



## **Abstract**

Replicability of research findings is a key issue in psychology, and has been the subject of considerable discussion by researchers. Replication is crucial to scientific principles, and underpins the consistency and verifiability of findings on which the foundation of scientific theories are built. Consistency of effects can be assessed through the perspective of heterogeneity, which indicates whether multiple research results into the same phenomenon are underpinned by the same true effect size. We use this perspective here to address concerns regarding replicability, and explore the application of heterogeneity in novel ways. This PhD project therefore aimed to: i) examine the heterogeneity of empirical findings in psychology; ii) consider the impact of biases in the research process on estimations of heterogeneity; iii) determine the strongest effects in psychology and their heterogeneity, and; iv) apply the perspective of heterogeneity to an existing psychological debate, the origin of psychological sex differences. In our first study, a re-analysis of 150 meta-analyses from cognitive, organisational, and social psychology was used to provide estimates of typical levels heterogeneity in psychology. This was compared to heterogeneity levels found across 57 multiple close replications. Our first study showed that low heterogeneity is achievable in psychology, but that typically heterogeneity estimates are high. Our next study used computer simulations to show that these levels of observed heterogeneity cannot be due to bias from questionable research practices or publication bias. In study 2, we showed the importance of including heterogeneity alongside effect size when interpreting an effect, and illustrated that even the largest mean effects in psychology might not be consistent in direction under conditions of high heterogeneity. Finally, in study 3 we showed that heterogeneity could provide an informative new perspective to the arguments regarding the origin of psychological sex differences. We explored several possible factors that could underlie heterogeneity in research findings, and suggest that unreliable forms of measurement and low quality of meta-analyses could provide promising avenues for further investigation in this regard. Overall, this thesis has provided an overview of the application of heterogeneity to provide a new perspective on psychological research.

# Contents

<b>Abstract</b> .....	<b>i</b>
<b>Contents</b> .....	<b>ii</b>
<b>List of Tables</b> .....	<b>viii</b>
<b>List of Figures</b> .....	<b>ix</b>
<b>Acknowledgements</b> .....	<b>x</b>
<b>Declaration</b> .....	<b>xi</b>
<b>1 Introduction</b> .....	<b>1</b>
1.1 Replication .....	1
1.2 Heterogeneity as a measure of consistency .....	2
1.3 Overview and objectives .....	2
1.4 Descriptive approach .....	3
<b>2 General methods</b> .....	<b>5</b>
2.1 Technical details.....	5
2.1.1 Effect sizes and conversion .....	5
2.1.2 Calculating standard deviations from standard errors .....	6
2.1.3 Choice of meta-analysis model .....	6
2.1.4 A note on terminology .....	6
2.1.5 Estimators of heterogeneity .....	7
2.1.5.1 Differences in the estimators.....	8
2.1.5.1.1 DerSimonian and Laird (DL estimator in Metafor) .....	8
2.1.5.1.2 Hunter and Schmidt (HS estimator in Metafor).....	8
2.1.5.1.3 Paule-Mandel (PM estimator in Metafor) .....	8
2.1.6 Calculation of variance.....	8
<b>3 Study 1a: Heterogeneity in the Results of Conceptual and Close Replications:</b>	
<b>Implications for Scientific Progress and Practical Applications</b> .....	<b>10</b>
3.1 Introduction .....	10
3.1.1 Measuring heterogeneity.....	12
3.1.2 Conceptual versus close replications .....	14

3.1.3	Moderators .....	15
3.1.4	Aims.....	16
3.2	Method .....	16
3.2.1	Study search and selection strategy .....	16
3.2.1.1	Inclusion criteria.....	16
3.2.2	Data extraction and analysis .....	32
3.3	Results.....	33
3.3.1	How meta-analyses address heterogeneity.....	33
3.3.2	Heterogeneity observed in close replications and conceptual replications.....	34
3.3.3	Moderators .....	35
3.3.4	Exploratory analyses on what drives heterogeneity.....	37
3.4	Discussion.....	39
3.4.1	Heterogeneity and the informativeness of meta-analyses.....	39
3.4.2	Replicability .....	40
3.4.3	Research planning.....	42
3.4.4	Generation of novel predictions .....	42
3.4.5	Empirical cumulativeness.....	43
3.4.6	Theoretical cumulativeness .....	45
3.4.6.1	Knowledge as a tool .....	45
3.4.6.2	Testing theories.....	46
3.4.7	Design of applications .....	46
3.4.8	Limitations.....	47
3.4.9	Conclusion and future directions.....	48
<b>4</b>	<b>Study 1b: Heterogeneity estimates and bias: A simulation study.....</b>	<b>50</b>
4.1	Introduction .....	50
4.1.1	Aims.....	52
4.2	Method .....	52
4.2.1	Factors manipulated .....	52
4.2.2	Technical details.....	54

4.3	Results.....	55
4.3.1	Choice of heterogeneity estimator .....	55
4.3.2	Bias in heterogeneity estimates.....	59
4.4	Discussion.....	61
<b>5</b>	<b>Study 2: Peaks of psychology: The very largest effects .....</b>	<b>63</b>
5.1	Introduction .....	63
5.2	Method .....	65
5.2.1	Study search and selection strategy .....	65
5.2.1.1	Inclusion criteria.....	65
5.2.2	Data extraction and analysis .....	67
5.2.3	Error and the meaning of very-large-effects .....	67
5.2.3.1	Assessment of the quality of meta-analyses .....	69
5.2.3.2	Assessment of triviality and informative nature of effects.....	70
5.3	Results.....	70
5.3.1	Our samples .....	70
5.3.2	How meta-analyses report heterogeneity.....	84
5.3.3	Are meta-analyses of very-large effects of particularly poor quality? .....	84
5.3.4	Number of studies in the meta-analysis .....	85
5.3.5	Meaning of meta-analyses.....	85
5.3.6	Exploratory analyses on what drives heterogeneity.....	86
5.4	Discussion.....	87
5.4.1	Peaks of psychology – the very largest effects .....	88
5.4.1.1	Clinical psychology: Treatment of mental health conditions .....	88
5.4.1.2	Cognitive psychology: Improving IQ and language, reading and communication skills in children.....	90
5.4.2	Error in meta-analyses .....	91
5.4.3	Distribution of large effects across sub-disciplines.....	91
5.4.4	Comparisons with the sample from study 1a .....	92
5.4.5	Limitations.....	92
5.4.6	Conclusion and future directions.....	93



<b>6</b>	<b>Study 3: Heterogeneity of sex differences across cultures .....</b>	<b>96</b>
6.1	Introduction .....	96
6.1.1	Evolutionary perspective .....	96
6.1.2	Social structuralist perspective .....	97
6.1.3	Why the consistency argument and the covariation argument are flawed. ....	98
6.1.4	Heterogeneity of sex differences.....	100
6.2	Method .....	101
6.2.1	Study search and selection strategy .....	101
6.2.1.1	Inclusion criteria.....	104
6.2.2	Data sources for sex differences.....	104
6.2.3	Data for psychological sex differences.....	105
6.2.3.1	Sex differences in mate preferences .....	105
6.2.3.2	Sex differences in mental rotation and line angle judgment.....	106
6.2.3.3	Sex differences in maths ability .....	109
6.2.3.3.1	PISA 2015 – age 15.....	109
6.2.3.3.2	TIMSS 2015 – 4 <sup>th</sup> grade .....	109
6.2.3.3.3	TIMSS 2015 – 8 <sup>th</sup> grade .....	109
6.2.4	Comparison data.....	112
6.2.4.1	Sex differences in body height.....	112
6.2.4.2	Sex difference in 2d:4d digit ratio.....	117
6.2.4.3	Cultural sex differences.....	120
6.2.5	Data extraction and analysis .....	124
6.3	Results.....	125
6.3.1	Comparison data.....	125
6.3.2	Psychological sex differences.....	128
6.3.3	Exploratory analyses .....	129
6.4	Discussion.....	129
6.4.1	Heterogeneity could be introduced through use of unreliable forms of measurement.....	130
6.4.2	Inferences regarding the origins of sex differences.....	131
6.4.3	Comparisons with the samples from study 1a and study 2.....	131

6.4.4	Strengths and limitations .....	131
6.4.5	Conclusions and future directions .....	132
<b>7</b>	<b>General Discussion.....</b>	<b>133</b>
7.1	Study 1a: Descriptive overview of heterogeneity.....	133
7.1.1	Implications of heterogeneity.....	134
7.2	Study 1b: Influence of bias: Questionable research practices and publication bias	134
7.3	Study 2: The largest effects in psychology.....	134
7.4	Study 3: Applying heterogeneity in a new context.....	135
7.4.1	Unreliable sources of measurement.....	136
7.5	Conclusions .....	136
	<b>Appendices .....</b>	<b>137</b>
	Appendix A.....	137
	Study 1a: References for all included meta-analyses and close replications .....	137
	Close replications .....	137
	Meta-analyses.....	137
	Cognitive .....	137
	Organisational.....	141
	Social .....	145
	Study 1a: Results based on Der Simonian Laird Hunter-Schmidt and Paule-Mandel estimators .....	149
	Descriptive statistics .....	149
	Robust t-tests for group differences.....	150
	Winsorized correlations between parameters .....	151
	Moderator analyses .....	152
	Median date split .....	152
	Predicting heterogeneity from effect size: Regression analyses .....	153
	Appendix B .....	154
	Study 1b: R code for simulations .....	154
	Study 1b: Results based on Der Simonian Laird, Hunter-Schmidt and Paule-Mandel estimators .....	180
	Study 1b: Interaction graphs for $T_{\text{bias}}$ based on DerSimonian-Laird, Hunter-Schmidt and Paule-Mandel estimators.....	181

Appendix C .....	182
Study 2: List of PsycINFO categories used .....	182
Study 2: Results based on Der Simonian Laird, Hunter-Schmidt and Paule-Mandel estimators .....	186
Study 2: Reason for classifying meta-analyses as ‘questionable’ in importance.....	190
Appendix D.....	193
Study 3: Reference list for height data .....	193
Study 3: Results based on Der Simonian Laird, Hunter-Schmidt and Paule-Mandel estimators .....	198
Study 3: Exploratory analysis of use of separate group means (N for each gender) versus total group size in meta-analysis .....	201
<b>References.....</b>	<b>202</b>

## List of Tables

Table 3.1 Meta-analyses in cognitive, organisational, and social psychology and close replication projects underpinning our heterogeneity analyses.....	19
Table 3.2 Descriptive Statistics for Study 1a.....	34
Table 4.1 Simulation Parameters for Study 1b. ....	54
Table 4.2 Initial effects of all simulation parameters on $T_{\text{bias}}$ . ....	61
Table 5.1 Number of meta-analyses for each sub-discipline (with percentage of sample) .....	73
Table 5.2 Means (and standard deviations) for meta-analysis quality and heterogeneity ( $T$ )....	73
Table 5.3 Correlations for exploratory analyses on what drives heterogeneity. ....	73
Table 5.4 Topics with the largest effect sizes in different sub-disciplines. Note that topics were grouped based on examination of similar effects. ....	74
Table 6.1 Sample sizes and effect sizes ( $d$ ) for sex differences on five aspects of mate preference.....	105
Table 6.2 Sample sizes and effect sizes ( $d$ ) for sex differences in mental rotation and line angle judgment tasks.....	107
Table 6.3 Sample sizes and effect sizes ( $d$ ) for sex differences on two measures of Maths ability .....	110
Table 6.4 Sample sizes and effect sizes ( $d$ ) for sex differences in height .....	113
Table 6.5 Sample sizes and effect sizes ( $d$ ) for 2d:4d digit ratio.....	118
Table 6.6 Sample sizes and effect sizes ( $d$ ) for sex differences in interest in science, and future career (interest in a career in STEM) .....	121
Table 6.7 Sample sizes and effect sizes ( $d$ ) for sex differences in attitudes toward gender roles and behaviours .....	123
Table 6.8 Descriptive statistics for effect size ( $d$ ) and heterogeneity ( $T$ ) of each sex difference .....	127

## List of Figures

Figure 3.1. Funnel plots for two meta-analyses. ....	11
Figure 3.2 Two distributions of population effect size (standardized mean differences).. ....	13
Figure 3.3 Sampling of meta-analyses. ....	18
Figure 3.4 Observed levels of heterogeneity for 57 close replications and for 50 meta-analyses (each) in cognitive, organisational and social psychology, with box-plots for $M_{win}$ . ....	35
Figure 3.5 Heterogeneity as a function of meta-analyses' mean effect size. ....	38
Figure 3.6 Ladder of virtue for independent variables/moderators. ....	44
Figure 4.1 $T_{bias}$ across all factor levels. ....	57
Figure 4.2 unsigned error in $T$ across all factor levels. ....	58
Figure 4.3 Distribution of $T_{bias}$ across all simulation conditions. ....	59
Figure 4.4 $T_{bias}$ for 1-tailed and 2-tailed testing, across different levels of effect size ....	60
Figure 5.1 Sampling of meta-analyses ....	66
Figure 5.2 Effects of error and bias on inclusion of meta-analyses in our sample.. ....	68
Figure 6.1 Two fictitious patterns of a sex difference across cultures. ....	99
Figure 6.2 Expected pattern of sex differences (based on fictitious data). ....	101
Figure 6.3 Relationship between observed sex differences and heterogeneity. ....	129

## Acknowledgements

Firstly, I would like to thank my principal supervisor, Dr Johannes Hönekopp for his guidance, advice, patience and support throughout the last 3 years. There's always time for another analysis, just for fun. I would also like to thank Dr Matthew Haigh for reading drafts and his comments and advice during write up.

I would like to thank Dr Thomas Pollet for his help with statistical analyses and R, and Dr Kris McCarty for lending multiple laptops for the simulation studies.

Next, I would like to thank the lovely people of CoCo for being such weird, fun, and supportive humans. I'm particularly thankful to Dr Anna-Marie Marshall for the long chats and support during meltdowns, Dr Léa Martinon for being a Word Ninja, and Claire McGrogan, Angela Wearn, Dimana Kardzhieva and Vicki Groves for helping me to keep going, general support and listening to my rants. Thanks also to Dale Metcalf for his eagle-eyed reference checking, and to Sarah Docherty and the CoCo games crew for much-needed distraction during write up.

I would like to thank Gabrielle Landric for all her care and support, and for believing in me. Thank you also to the SABBS and staff of Northumbria Students' Union for the opportunities you gave me, and the support with difficult circumstances.

I would also like to thank my academic mum, Emma Jones-Holding, for the very long phone calls, sanity checks, Netflix access, and for helping me figure myself out and keep going. Still an academic parent all these years later!

Finally, I would like to thank my wife Katie Linden for being incredibly supportive, patient and tolerant of the whole PhD process. Thank you for not giving up on me and believing I could do it. I wouldn't be here without you.

## **Declaration**

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research in this thesis has been approved. Approval has been sought and granted by the Faculty Ethics Committee on 08/11/2016 (Study 1, Ref: SUB008\_Linden\_081116), 07/12/2017 (Study 2, Ref: 4274), and 13/09/2018 (Study 3, Ref: 10892).

I declare that the word count of this thesis is: 41,623 words

Name:

Signature:

Date:

# **1 Introduction**

The issue of replicability of research results in psychology has been the subject of much recent debate and concern within the discipline. This is perhaps most commonly illustrated through a recent attempt to directly replicate the findings of 100 experimental and correlational studies (Open Science Collaboration, 2015). In this replication attempt, the Open Science Collaboration (2015) suggest that replication effects were, on average, only half the magnitude of effects found by the original studies, and that replication attempts were successful in only 39% of cases.

A single, false-positive result (type-1 error) can occur due to chance caused by sampling error, that is, error caused by using a sample that is unrepresentative of the population one is intending to study. On its own, this does not have to be a problem for scientific progress as long as scientific processes of self-correction remain in place – the type-1 error will be uncovered and the record can be corrected (Ioannidis, 2012). However, replication is essential to separate these false alarms from real discoveries. Reproducibility of findings is thus key to scientific principles – observations should be able to be replicated by anyone who conducts research according to the same methodology, as the verifiability of observations by independent sources underpins the notion of what constitutes science (Popper, 1959).

## **1.1 Replication**

In brief, replication refers to attempts to reproduce an observed effect, either through observation or through intervention. Close replications aim to repeat an earlier study as closely as possible, whereas conceptual replications address the same phenomenon but deviate from the original study in key ways, for example aspects of study design, materials, participants or analysis.

The interpretation of replication studies in psychology has created extensive debate, which we will not duplicate here (see Zwaan, Etz, Lucas, & Donnellan, 2018, for a comprehensive summary). However, a number of challenges and objections to the value of replications are worth considering. Firstly, the context of the original study (for example geographical, cultural or historical) could have changed over time, which could affect the replicability of the result itself. It is possible that seemingly irrelevant features of the context, or of the study itself, which may not have been reported, could have affected the study in unanticipated ways, and



render an effect impossible to replicate. These seemingly irrelevant features are sometimes referred to as hidden moderators of the effect, and could be used to explain some instances of replication failure (Baribault et al., 2017). However, it is arguable that such effects could instead be false positives, and fail to replicate for this reason. If an effect is so specific to the context in which it was conducted, it is also arguable how much this result is worth – hypotheses are based on predictions about expected outcomes, and if these predictions do not include context-specific details, the hypothesis should hold under a variety of conditions. In addition to this, it is also important to assess under which conditions an effect holds, and replication attempts can therefore provide important value in this respect.

Additionally, some argue that replication failure is not a serious issue facing psychology as a discipline because meta-analyses in most of its research areas show that we already have meaningful replication of phenomena, which can provide answers to research questions (Schmidt & Oh, 2016)<sup>1</sup>. However, for this argument to hold, it is important to consider the consistency of research findings within meta-analyses. This is the aim of Study 1, which considers studies in a meta-analysis as replications of the same phenomenon and asks how coherent or incoherent their results typically are. As we will discuss in Study 1, the situation is not as straightforward as Schmidt and Oh imply.

## **1.2 Heterogeneity as a measure of consistency**

Meta-analysis can provide an indication of consistency of effects by examining the heterogeneity of findings. Heterogeneity indicates whether the studies summarised in the meta-analysis tap into the same true effect size: if this were the case, true heterogeneity would be zero. More formally defined, heterogeneity relates to the amount of variability observed in effect sizes over and above that which may be expected due to sampling error, that is, error attributable to the fact that a study sample may not be representative of the population it is intended to represent.

## **1.3 Overview and objectives**

Study 1a aims to examine the heterogeneity of empirical findings in psychology across meta-analyses published in a variety of journals, and provide an indication of the replicability of

---

<sup>1</sup> Briefly, the technique of meta-analysis strives to summarise the results of multiple studies into the same phenomenon. See Study 1a for more detail concerning this method.

research. From this descriptive work, a number of important questions and implications arise. Firstly: what is the impact of biases in the research process on estimations of heterogeneity? This is addressed in Study 1b. Second, all else being equal, strong effects should be of particular interest in understanding and solving problems in psychology, particularly if they show little heterogeneity: what are the strongest effects in psychology, and which areas of psychology do these refer to? This is addressed in Study 2. Finally, can we use the perspective of heterogeneity to attempt to provide a new perspective on an existing debate in psychological research, namely the underpinning of psychological sex differences? This is addressed in Study 3.

Most research in this thesis is descriptive in nature, and this is sometimes frowned upon in psychology where theory testing is often seen as the only route to progress. We therefore close our introduction with a brief justification of our descriptive approach.

#### **1.4 Descriptive approach**

Research in psychology has predominantly focused on theory development and hypothesis testing rather than descriptive work (Greenwald, 2012). Unfortunately, this has led to a situation where many theoretical controversies persist, and theories are difficult to resolve or disprove, perhaps especially because crucial findings can be dismissed when they oppose a prominent theory, and are treated as conceptually or methodologically flawed (Greenwald, 2012). This can even be the case when studies are combined using meta-analysis, where several meta-analyses can be conducted into the same phenomenon, by authors with different theoretical standpoints, that find apparently contradictory results (Ferguson & Heene, 2012). However, description of phenomena should precede, and form the basis of, theory development, in order that properties and their generality can be established, and appropriate methodologies determined, before proposing and testing theories (Rozin, 2009). As perhaps more mature disciplines have illustrated (Fanelli, 2010; Fanelli & Ioannidis, 2013), failure to do so can lead to fundamental concerns about bias, accuracy of effect size estimates, replicability, and failures in the normally self-corrective process of science (Ferguson, 2015b; Ioannidis, 2012; Schimmack, 2012).

There are numerous examples of key contributions to scientific discovery that have stemmed from such descriptive work. To take an example from Earth sciences, theories of plate tectonics were developed from taking an overview of geologic features across continents. By taking a broader view, the fit of shorelines between different continents was observed,

patterns of fossils and rock formations were highlighted, and the movement of continents was suggested as underpinning the patterns observed. This informed proposal of the theory of thermal conduction as the mechanism underpinning plate tectonics, which was later developed and confirmed by discovery of further evidence to support this, such as discovery of mid-oceanic ridges and oceanic trenches at continental margins (Grotzinger, Jordan, & Press, 2014). Taking a descriptive overview of empirical results of research within a field can therefore illustrate broader patterns of findings and illuminate issues within the discipline, as well as informing the direction of future research (Rozin, 2009).

In psychology, a limited number of descriptive overviews of specific research areas have been undertaken in the form of meta-syntheses of meta-analyses (for example, Johnson, Scott-Sheldon, & Carey, 2010; Lipsey & Wilson, 1993; Richard, Bond, & Stokes-Zoota, 2003; Zell, Krizan, & Teeter, 2015), and examination of methodological concerns (Levine, Asada, & Carpenter, 2009; Stanley, Carter, & Doucouliagos, 2018; Van Erp, Verhagen, Grasman, & Wagenmakers, 2017). We aim to add to this contribution.

## 2 General methods

The purpose of this chapter is to provide an overview of the general methodology used throughout this thesis. Here we provide an overview of technical details including use and conversion of effect sizes, calculation of standard deviations, choice of meta-analytic model and the estimators of heterogeneity used throughout all studies in the thesis.

### 2.1 Technical details

#### 2.1.1 Effect sizes and conversion

We chose Cohen's  $d$  (Standardised Mean Difference) as our effect size throughout this thesis, as this is perhaps the most commonly used standardised effect size measure in psychology (Fanelli, Costas, & Ioannidis, 2017). Where meta-analyses utilised correlational studies, we converted these to Cohen's  $d$  following Borenstein, Hedges, Higgins, and Rothstein (2009) using the following formula:

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

Equation 2.1

The close relationship between the concepts of correlations and standardised differences, and the ease of conversion from one to the other, determined our choice to include meta-analyses utilising both of these measures of effect size. Conversions between other types of effects that use categorical dependent variables are less conceptually sensible (Ferguson, 2009).

Where these proved unavoidable (study 2), we converted Cohen's  $d$  into log odds ratio (Log OR) using the following formula, following Borenstein, Hedges, Higgins, and Rothstein (2009), (where  $\pi$  denotes the mathematical constant, approximately 3.142):

$$\text{Log Odds ratio} = d \frac{\pi}{\sqrt{3}}$$

Equation 2.2

For the same purpose, we converted Log OR to OR. Log OR is the natural logarithm of OR (represented mathematically as  $\ln \text{OR}$ , or as  $\log_e \text{OR}$ ). Therefore, to convert from Log OR to OR, the inverse logarithm is used ( $e^{\log \text{OR}}$ ), where  $e$  is the mathematical constant that is the base of the natural logarithm (approximately 2.718).

### 2.1.2 Calculating standard deviations from standard errors

Where studies failed to report required standard deviations, we computed the latter from standard errors as

$$SD = SE \sqrt{n}$$

Equation 2.3

where  $n$  is the sample size.

### 2.1.3 Choice of meta-analysis model

Meta-analysis seeks to combine the results of multiple studies into the same phenomenon. Often the main goal of meta-analysis is to estimate the overall effect relating to that phenomenon. Broadly, studies are weighted so that more precise estimates of an effect (gained from studies with a larger sample size) are assigned greater weight than smaller studies; however, the exact method for weighting studies can vary depending on the model of meta-analysis used.

There are two main forms of models used in meta-analysis, the fixed effects model, and random effects model. The fixed effects model assumes that there is a single population effect size underpinning all studies, whereas the random effects model assumes there is not a single population effect size underpinning the included studies, and the 'true' effect might vary. In this latter case, it is usually assumed that the population effects follow a normal distribution, and studies in the meta-analysis are a sample from this distribution. The standard deviation of this distribution is  $\tau$  (Tau), and this is a measure of heterogeneity, that is, the variability in effect sizes over and above that expected due to sampling error.

All meta-analyses conducted in this research have involved use of the random effects model. Several different estimators of heterogeneity have been developed within this model, and some of these will now be covered in brief.

### 2.1.4 A note on terminology

For Cohen's  $d$ ,  $\delta$  refers to the population effect size, and  $d$  refers to the estimate of this parameter. Similarly for heterogeneity,  $\tau$  refers to the standard deviation of the population of effect sizes, and  $T$  refers to the estimate.

### 2.1.5 Estimators of heterogeneity

For study 1a, and study 2, we recomputed meta-analyses based on the data reported in the published meta-analyses. This data included an effect size and sample size for each of the primary studies that the original authors included in their meta-analysis. For some meta-analyses, means and standard deviations were available, but we used Cohen's  $d$  in these cases to ensure consistency in calculations for all meta-analyses in our sample. This was particularly relevant as some of the original meta-analyses included in study 1a or study 2 were based on correlations, which we converted to Cohen's  $d$ . This meant that we had to calculate the variance based on an assumption of equal groups. For study 3, we were able to obtain raw data for each of the groups for the primary studies included. This allowed us to use the means and standard deviations as well as sample sizes for each study, which provides a more accurate calculation of variance for use in the meta-analysis where group sizes are unequal. In addition to this, all meta-analyses in study 3 were conducted on sex differences, therefore all studies were between groups designs.

For all studies, mean effect size (Cohen's  $d$ ) and heterogeneity ( $T$ ) for each meta-analysis was calculated using the package Metafor in R (Viechtbauer, 2010). Data collection for a pilot version of study 2 conducted alongside study 1a included recording the meta-analysis model and specific estimator utilised by the original authors of each meta-analysis. This allowed us to determine the frequency of use of different estimators of  $T$ . The Hunter Schmidt method was used most frequently (in 41 meta-analyses, 30 of which were in organisational psychology), followed by DerSimonian and Laird (in 22 meta-analyses. Note that this is the estimator discussed by Borenstein, Hedges, Higgins, and Rothstein (2009) in their popularly used textbook introduction to meta-analysis).

However, the DerSimonian and Laird approach can be negatively biased when heterogeneity is moderate to high (Langan, Higgins, & Simmonds, 2015; Langan, Higgins, & Simmonds, 2017; van Aert & Jackson, 2018; Veroniki et al., 2016). Several authors now recommend the use of Paule-Mandel estimator of between-study variance for random effects meta-analyses (for example, Langan et al., 2015; Langan et al., 2017; van Aert & Jackson, 2018; Veroniki et al., 2016) as this estimator does not suffer from the problem of negative bias in heterogeneity, and is both relatively simple to implement and available in Metafor. We therefore utilised three estimators of heterogeneity in all studies: DerSimonian and Laird, and Hunter Schmidt, due to their popularity in psychology; and Paule-Mandel as a recommended alternative.

### 2.1.5.1 *Differences in the estimators*

Differences between the estimators of heterogeneity predominantly arise from differences in the method of estimating the sampling error variance of the mean effect size.

#### 2.1.5.1.1 DerSimonian and Laird (DL estimator in Metafor)

This is the most popular estimator, perhaps because it is straightforward, and is recommended in a popular introductory textbook (Borenstein et al., 2009). In this case heterogeneity is calculated from  $T^2$ , using the method of moments method (see Borenstein et al., 2009, p72-74 for details). This estimator is also sometimes referred to an inverse variance method, as it weights each study using the inverse of within-study variance plus between studies variance ( $T^2$ ), and produces a weighted mean effect size based on the sum of the products (effect size x weight) divided by the sum of the weights.

#### 2.1.5.1.2 Hunter and Schmidt (HS estimator in Metafor)

The Hunter and Schmidt estimator is particularly popular in organisational psychology, where psychometric meta-analyses and corrections for artifacts (such as error of measurement and range variation) are common (Hunter & Schmidt, 2004). This estimator uses the variance of the observed effect sizes ( $ds$ ) across studies, divided by the number of studies ( $k$ ) to estimate the sampling error variance of the mean effect size. The study weights used in this model are based on total sample sizes, and this could therefore be expected to be a less accurate method when group sizes for experimental and control groups are unequal.

#### 2.1.5.1.3 Paule-Mandel (PM estimator in Metafor)

We included this estimator based on the recommendation of several authors who state that this is a useful alternative to the DerSimonian and Laird (DL) estimator, as it avoids the negative bias that the DL estimator can suffer from under conditions of moderate to high heterogeneity (Langan et al., 2015; Langan et al., 2017; van Aert & Jackson, 2018; Veroniki et al., 2016).

### 2.1.6 Calculation of variance

For study 1a and study 2, we imported the sample sizes and Cohen's  $d$  for each meta-analysis into Metafor, and calculated the sampling error variance of the mean effect size in R using a simple calculation based on an assumption of equal groups. For study 3, effect sizes (Cohen's  $d$ ) and variance were computed using the *escalc* function (part of the Metafor package) in R.

This latter approach provides a more accurate estimate of variance when group sizes are unequal, and was possible in this case as all studies involved between-groups designs examining sex differences.



### **3 Study 1a: Heterogeneity in the Results of Conceptual and Close Replications: Implications for Scientific Progress and Practical Applications**

#### **3.1 Introduction**

Meta-analysis, which seeks to summarize results of multiple studies into the same phenomenon, has become an indispensable tool in contemporary research. In pioneering work, Smith and Glass (1977) showed that psychotherapy has a strong positive effect on the average patient studied, and Schmidt and Hunter (1977) demonstrated that the validity of employment tests generalise more readily across different job types than previously believed. Roughly, the result of a meta-analysis addresses two inferences: the mean effect size in a population of studies, and their heterogeneity. The latter indicates if the studies summarized in the meta-analysis tap into the same true effect size. In this case, heterogeneity would be zero. Sampling error should always create some differences in observed effects across studies. Zero heterogeneity is inferred where these observed differences do not exceed the level expected due to sampling error. Consider the effectiveness of psychotherapy as an example. If heterogeneity was zero, the effectiveness of psychotherapy would be the same across all studies, regardless of the issue patients present with (e.g. anorexia, depression, specific phobia), the type of therapy they receive (e.g. cognitive-behavioural, psychoanalytic), and other differences. Obviously, this is unrealistic; for example, some conditions are treated more successfully than others (Huhn et al., 2014). Heterogeneity then reflects how much the true effect sizes differ across studies. Figure 3.1 provides examples with high and low heterogeneity.

Meta-analyses tend to focus more on mean effect size than heterogeneity (Ioannidis, 2008). For example, virtually all of them emphasize if the mean effect size differs significantly from zero. In contrast, heterogeneity tends to be neglected in psychological meta-analyses (Dieckmann, Malle, & Bodner, 2009). Recently however, a number of systematic investigations started to shift heterogeneity's Cinderella status: Several demonstrated that heterogeneity typically decreases the statistical power of studies<sup>2</sup>, i.e. any real effect under investigation is less likely to produce a statistically significant result (Kenny & Judd, 2019; McShane &

---

<sup>2</sup> This is because the gain in power for larger-than-expected effects is less than the loss in power for smaller-than-expected effects.

Böckenholt, 2014; Shrout & Rodgers, 2018), and a systematic survey found that heterogeneity tends to be large in psychology (Van Erp et al., 2017). Stanley, Carter, and Doucouliagos (2018) combined both approaches to estimate the power of replication studies and found that this will be typically very low, including close replications (e.g. Open Science Collaboration, 2015). Here, we re-examine observed levels of heterogeneity and its consequences. We describe heterogeneity in meta-analyses and in close replication studies and we investigate if moderator variables can successfully explain heterogeneity, which would suggest that previous accounts (Stanley et al., 2018; Van Erp et al., 2017) exaggerated the latter. To foreshadow our results, we confirmed that heterogeneity tends to be large outside close replications and found the effects of moderators to be negligible; however in stark contrast to Stanley et al.'s findings, we found heterogeneity to be negligible in close replications. In our discussion, we move beyond the important but narrow issue of statistical power and explore the wide-ranging implications of this finding for progress in psychological science and the latter's successful application to practical problems.

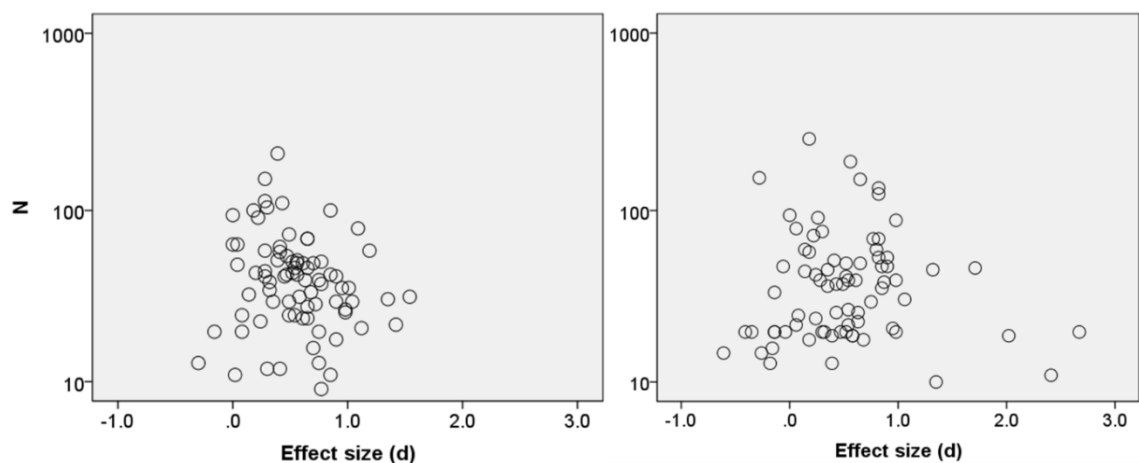


Figure 3.1. Funnel plots for two meta-analyses

Note. *Left-hand panel*: Low heterogeneity. Linck, Osthus, Koeth, and Bunting (2014), investigated the link between working memory and second language comprehension. Estimated mean of the population of effect sizes  $d = 0.51$ , standard deviation of observed effect sizes is 0.36, estimated heterogeneity of true effect sizes  $T = 0.11$ ,  $I^2 = 11$ . *Right-hand panel*: High heterogeneity. Baker, Peterson, Pulos, and Kirkland (2014), investigated the link between intelligence and performance on the Reading the Mind in the Eyes test.  $d = 0.49$ ,  $SD = 0.59$ ,  $T = 0.35$ ,  $I^2 = 53$ .

### 3.1.1 Measuring heterogeneity

Before we describe substantive issues like type of replication, moderators, and biases in greater detail, it is helpful to address how heterogeneity can be quantified. In psychology the idea of heterogeneity is usually discussed in the context of standardized effect sizes (e.g. Cohen's  $d$  instead of a difference between group means in raw scores), and we stick to this perspective here. Three established approaches to deal with heterogeneity are Cochrane's  $Q$  statistic,  $I^2$ , and  $\tau$ . Before we deal with them, it is helpful to consider what we should expect to see in the absence of heterogeneity. Even if all primary studies tap into the same population effect size, we expect to see differences in the observed effect sizes due to sampling error. Thus, observed differences between effect sizes do not necessarily point to heterogeneity. One of the most popular approaches (Cochrane's  $Q$ ) computes how likely the observed differences between effect sizes would be if all studies tapped into the same population effect size. It thus tests if observed effect size differences are significantly larger than expected under the null hypothesis of zero heterogeneity (Higgins, Thompson, Deeks, & Altman, 2003). This approach does not quantify heterogeneity, and is underpowered for meta-analyses with few studies.

Another approach,  $I^2$ , quantifies heterogeneity. It reflects the percentage of observed effect size variability that reflects real differences in effect sizes. Where  $I^2$  is near zero, the observed variability is mostly down to sampling error; where  $I^2$  is near 100, most of the observed variability reflects differences in population effect sizes. Unfortunately,  $I^2$  depends on the sample size in the primary studies (Borenstein, Higgins, Hedges, & Rothstein, 2017; IntHout, Ioannidis, Borm, & Goeman, 2015). Imagine that all studies used small samples. Individual effect sizes will scatter widely around their true effect size. Consequently, a large percentage of observed variability reflects sampling error and  $I^2$  will be low. Now imagine that all studies used very large samples. Each study will provide a highly accurate estimate of its population effect size. Consequently, only a small percentage of observed variability reflects sampling error and  $I^2$  will be high. Therefore, comparing meta-analyses on  $I^2$  strikes us as questionable unless average sample sizes are similar.

The third approach directly estimates the variability in population effect sizes. It is generally assumed that population effect sizes relating to a given phenomenon follow a normal distribution;  $\tau$  refers to their standard deviation (Borenstein et al., 2009) As an example, consider the meta-analysis in the left panel of Figure 3.1. The standard deviation of the observed effect sizes in the primary studies is 0.36. (For the sake of consistency, we use

Cohen's  $d$  as a measure of effect size in this example and throughout our paper.) Some of the observed effect size variability must be down to sampling error. When this is removed, heterogeneity is estimated as only 0.11. To better understand heterogeneity, consider Figure 3.2. Here, the mean for the population of true effect sizes is  $\delta = 0.45$  and their standard deviation is 0.33, therefore  $\tau = 0.33$ . The standard deviation roughly reflects how far data points are, on average, away from the mean. Consequently, any study's population effect size will typically deviate from 0.45 by approximately 0.33. As  $\tau$  quantifies heterogeneity and lends itself to comparisons across meta-analyses, we use it here<sup>3</sup>. As with Cohen's  $d$  – where  $\delta$  refers to the population effect size, and  $d$  refers to the estimate of this parameter –  $\tau$  refers to the population SD, and  $T$  refers to its estimate. For any meta-analysis, the precision of  $T$  as an estimate for  $\tau$  increases with the number of studies ( $k$ ).

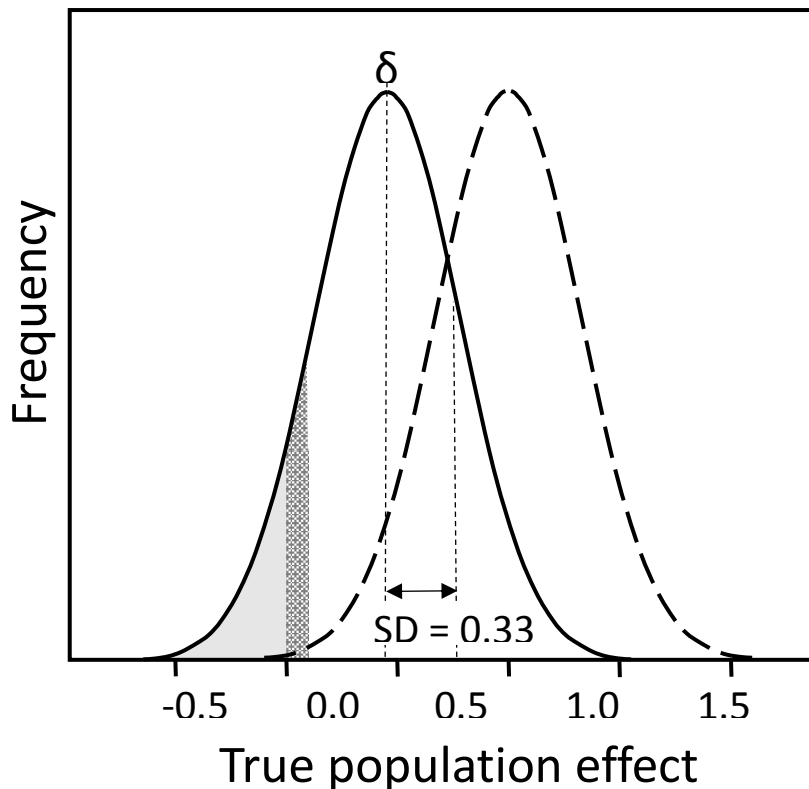


Figure 3.2 Two distributions of population effect size (standardized mean differences).

*Note.* The distribution on the left (solid line) shows a population effect size with a mean ( $\delta$ ) of 0.45 and a standard deviation ( $T$ ) of 0.33. 9% of true effect sizes are opposite of the expected direction ( $FI$ , light grey area); an additional 5% of true effect sizes are in the expected direction but very small ( $< 0.1$ , dark grey area). The distribution on the right (dashed line) shows a population effect size with a mean ( $\delta$ ) of 1 and a standard deviation ( $T$ ) of 0.33.

<sup>3</sup> Indeed, across 141 meta-analyses based on Cohen's  $d$ ,  $I^2$  and  $T$  correlated only moderately at  $r = .39$  (Van Erp et al., 2017).

The same level of heterogeneity has different implications depending on the mean effect size. Consider a beneficial treatment with an average true effect size of  $\delta = 1$  and heterogeneity of  $\tau = 0.33$  (Figure 3.2, dashed curve). This means that the true effectiveness of the treatment differs considerably across studies, however, under virtually all circumstances the effect remains positive. Now contrast this with a beneficial treatment that shows the same heterogeneity, but has a lower average true effect size (Figure 3.2, solid line). The effectiveness of the treatment is equally variable, however for a substantial proportion of studies, the treatment now proves harmful (i.e. the true effect size is less than zero). We therefore find it useful to also consider the proportion of true effect sizes that go against the expected direction. We introduce a new way of referring to this proportion, namely the flop index (*FI*). This is indicated by the light grey area under the curve in Figure 3.2. Note that our conceptualisation of the *FI* is conservative. In many research contexts, positive but small true effects are not useful. Consider true effects of  $0 < \delta \leq 0.1$  (dark grey area in Figure 3.2). In a between-subjects study, we would need at least 1,570 participants per condition to have sufficient statistical power (assuming typical parameters of 80% power,  $\alpha = 0.05$ , and 2-tailed testing); for  $\delta = 0.2$ , this number is still 390 per condition. In many research contexts, such sample sizes are unrealistic. Where our conservative *FI* in Figure 3.2 is 9%, it would rise to 14% and 22% for the proportion of studies that tap into true effects of  $\delta < 0.1$  and  $\delta < 0.2$  respectively. In some applied contexts however, even very small beneficial effects are worthwhile (Rosenthal & DiMatteo, 2001); therefore, our *FI* considers only true effects that run against the expected direction as flops.

### 3.1.2 Conceptual versus close replications

For any discussion of the heterogeneity of research findings and its implications, it is essential to differentiate between conceptual and close replications (Schmidt, 2009; Zwaan et al., 2018). The latter seek to replicate an earlier study as faithfully as possible. The Open Science Collaboration (2015) project is a famous example. In a massive collaborative effort, the authors sought to replicate 100 studies recently published in high profile psychology journals. The replications sought to copy study materials, data analyses, and other key aspects of the original studies as closely as possible. Although the replications were thought to have high power, only 36% achieved statistically significant results. However, power calculations did not take heterogeneity into account. Stanley et al. (2018) argued that heterogeneity in close replications is high and that the low replication rate in the OSF is readily explained by the resulting low statistical power. Critically, their estimate of heterogeneity in close replication

studies ( $T = 0.25$ ) was based on only two studies (Eerland et al., 2016; Hagger et al., 2016). Note that the heterogeneity for each of the 100 twin studies in the OSF (original and replication) cannot be estimated reliably. Instead of a single replication, this would require multiple close replications of the same effect (hence, Many-Labs type replications, Klein et al., 2014). Here, we analyse 57 Many-Labs type replications in order to derive a more useful estimate of heterogeneity in close replications.

The studies summarized in a meta-analysis can typically be considered to be conceptual replications (Schmidt & Oh, 2016), i.e. whilst they address the same topic or mechanism, they often differ markedly regarding their design, study materials, participants, data analysis, and other key aspects. Consequently, heterogeneity should tend to be larger in conceptual replications than in close replications. Previous heterogeneity surveys were based on meta-analyses in Psychological Bulletin (Stanley et al., 2018; Van Erp et al., 2017) and found high levels of heterogeneity (median  $T$  of 0.35 in Stanley et al.). We decided to look at meta-analyses from cognitive, organisational, and social psychology because these sub-disciplines received particular attention in previous meta-scientific investigations (Mitchell, 2012; Open Science Collaboration, 2015). Mitchell (2012) compared effect sizes from laboratory-based and field-based studies in organisational and social psychology. The correlation between lab- and field-based effect sizes was higher in the former ( $r = 0.89$ ) than in the latter ( $r = 0.53$ ). Open Science Collaboration found that cognitive psychology findings replicated substantially better than social psychology findings. These observations could point to greater heterogeneity within social psychology in general. Indeed, Stanley et al. (2018) argued that the low replication rate observed in the Open Science Collaboration might reflect low power caused by high heterogeneity. From this perspective, the difference in observed Open Science Collaboration replication rates would then suggest higher heterogeneity in social psychology as compared to cognitive psychology. We test this idea here.

### 3.1.3 Moderators

Previous surveys of heterogeneity (Stanley et al., 2018; Van Erp et al., 2017) considered each meta-analysis at the highest level of aggregation. At this, apples and oranges might be unwisely mixed (thus artificially inflating heterogeneity), where instead separate analyses on apples and oranges would be in order. Where heterogeneity is inflated by mixing apples and oranges, a powerful moderator (type of fruit) should emerge. Once this is used to split the studies into sensible subsets, heterogeneity within each group should be lower. Here we

looked into the effectiveness of moderators to decrease heterogeneity by also analysing the latter in moderator-determined subsets of effects.

#### 3.1.4 Aims

In sum then, our goals are: i) to derive a reliable estimate of the heterogeneity in close replications and to compare this against levels of heterogeneity in conceptual replications; ii) to investigate if heterogeneity is particularly high in social psychology, thus reflecting the low replication rate for this field in Open Science Collaboration (2015); iii) to investigate to what extent moderators can explain heterogeneity. We use these findings to discuss implications for replicability, research planning, theory evaluation, and the design of practical applications.

### 3.2 Method

#### 3.2.1 Study search and selection strategy

We intended to investigate all available Many-Labs type replications. We searched Curate Science (2017) for relevant reports in April 2017 and added studies from Many Labs 2 (Klein et al., 2018) at a later stage. Further, we thought to investigate 50 meta-analyses each from cognitive, organisational, and social psychology. Feasibility, rather than power considerations, determined this choice. Our pre-registered study protocol is available at:

<http://aspredicted.org/blind.php?x=bf46k8>.

In order to obtain our sample, we searched PsycINFO, journals only, for “meta-analy\*” in the abstract field, in November 2016. We restricted searches to PsycINFO classifications “3000 Social Psychology”, “3600 Industrial and Organisational Psychology”, and for cognitive psychology “2340 Cognitive Processes”, “2343 Learning and Memory”, and “2346 Attention”. Not enough eligible meta-analyses could be obtained in this way (see below for inclusion criteria). We therefore searched Web of Science, articles only, for “meta-analy\*” in the categories “Psychology Social”, “Psychology Applied”, and “Psychology”, excluding meta-analyses that fell outside our target sub-disciplines (see Figure 3.3). These were inspected in random order until we reached the desired number of 50 meta-analyses.

##### 3.2.1.1 *Inclusion criteria*

For the three sub-disciplines, meta-analyses were included if they met all of the following criteria. i) It addressed a substantive psychological effect (rather than, for example, the

psychometric properties of a questionnaire). ii) The analysed effects were described as standardized mean differences (Cohen's  $d$ , Hedge's  $g$ ) or correlations (Pearson's  $r$ , or Fisher's  $Z$ ). Standardized differences and correlations are closely related concepts, and one can easily be converted into the other. Similar conversion are less sensible were categorical dependent variables are used (Ferguson, 2009), and our heterogeneity measure  $T$  is also not suitable for these types of effect sizes. For this reason, we excluded meta-analyses that used odds ratios, risk ratios, and similar measures. iii) Effect size and sample size information was provided for the original studies. This was necessary to calculate heterogeneity. Where only the sample sizes or effect sizes were available, an attempt was made to obtain missing data from the corresponding author of the meta-analysis. iv) For practical reasons, the full article had to be available in English. All close replication reports that met the same criteria (Many Labs 1, 2 and 3, and Registered Replication Reports 3 to 6) were included (Cheung et al., 2016; Ebersole et al., 2016; Eerland et al., 2016; Hagger et al., 2016; Klein et al., 2014; Klein et al., 2018; Wagenmakers et al., 2016)<sup>4</sup>.

In this way, we identified 50 meta-analyses for cognitive, organisational, and social psychology each, and 57 for close replications (see Table 3.1).

---

<sup>4</sup> For Many Labs 2 (Klein et al., 2018), effect sizes for individual studies were not reported, but we could compute them from the published raw data.



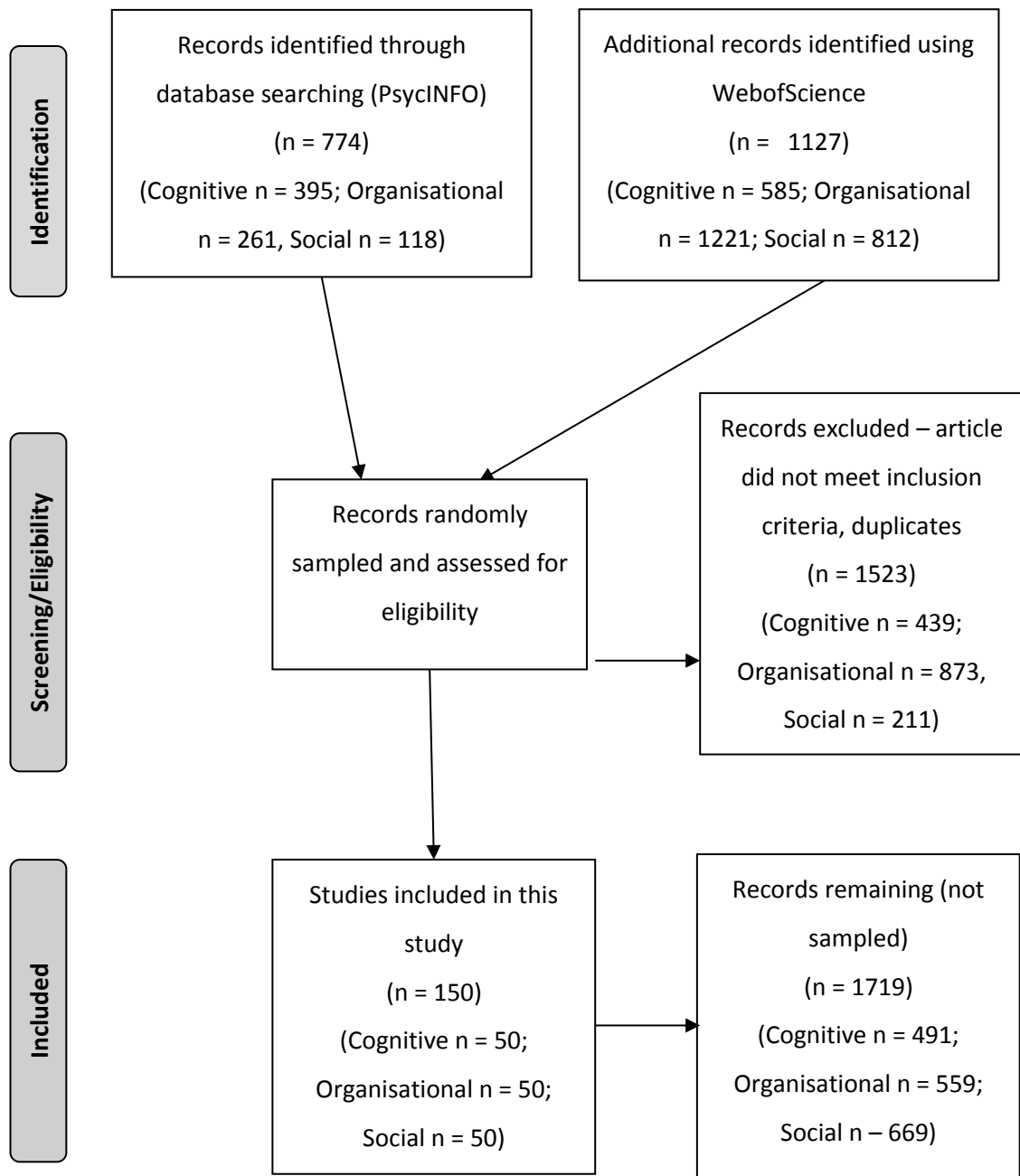


Figure 3.3 Sampling of meta-analyses.

Table 3.1 Meta-analyses in cognitive, organisational, and social psychology and close replication projects underpinning our heterogeneity analyses. See Appendix A for full references.

Reference	Topic
<b>Cognitive Psychology</b>	
Adesope, Lavin, Thompson & Ungerleider (2010)	Cognitive correlates of bilingualism
Allen, Berkowitz, Hunt & Louden (1999)	Impact of forensics and communication education on critical thinking
Alvarez & Emory (2006)	Relationship of executive function to frontal lobes
Au, Sheehan, Tsai, Duncan, Buschkuehl & Jaeggi (2015)	Improving fluid intelligence with working memory training
Baker, Peterson, Pulos & Kirkland (2014)	Relationship between Intelligence and “Reading the Mind in the Eyes”
Beaujean (2005)	Heritability of cognitive abilities
Bodner, Taikh & Fawcett (2014)	Assessing the costs and benefits of producing words aloud in recognition
Chen & Li (2014)	Association between non-symbolic number acuity and math performance
D'Alessio & Allen (2002)	Association between selective exposure and cognitive dissonance after decisions
Daneman & Merikle (1996)	Association between working memory and language comprehension
Dato-on & Dahlstrom (2003)	Influence of context and contrast on judgments in decision making
Dewald, Meijer, Oort, Kerkhof & Bögels (2010)	Influence of sleep on school performance
Doyle & Voyer (2016)	Stereotype manipulation effects on math and spatial test performance
Eastvold, Belanger & Vanderploeg (2012)	Effects of third party observer on neuropsychological test performance
Evers & Lakens (2014)	Tversky’s diagnosticity principle – which features are used to cluster objects into subgroups?

Fawcett (2013)	The production effect benefits performance
Herlitz & Loven (2013)	Sex differences and own-gender bias in face recognition
Hinshaw (1991)	Effects of mental practice on motor skill performance
Huff, Bodner & Fawcett (2015)	Effects of distinctive encoding on correct and false memory
Linck, Osthus, Koeth & Bunting (2014)	Relationship between working memory and second language comprehension and production
Lucas (1999)	Effect of context on lexical access
Maeda & Yoon (2013)	Gender differences in mental rotation
McDaniel (2005)	Relationship between brain volume and intelligence
McMorris & Hale (2012)	Effects of acute exercise on speed and accuracy of cognition
Meijer, Selle, Elber & Ben-Shakhar (2014)	Memory detection with the Concealed Information Test
Merikle & Daneman (1996)	Memory for unconsciously perceived events
Moxley & Charness (2013)	Age and skill effects on recalling chess positions and selecting best move
Nieuwenstein, Wierenga, Morey, Wicherts, Blom, Wagenmakers & van Rijn (2015)	Unconscious thought advantage – does performing an unrelated task affect performance on a subsequent task
Park (2005)	Effect of arousal and retention delay on memory
Pässler, Beinicke & Hell (2015)	Relationship between vocational interests and intelligence
Pietschnig, Voracek & Formann (2010)	Mozart effect – effect of listening to Mozart on performance on spatial tasks
Powers, Brooks, Aldrich, Palladino & Alfieri (2013)	Effects of video-game play on information processing
Preiss, Wheelless & Allen (1990)	Relationship between receiver apprehension and cognitive processes

Redick & Lindsey (2013)	Relationship between complex span and n-back measures of working memory
Reinwein (2012)	Does the modality effect exist?
Rice & Mullen (2003)	Determining the accuracy of ethnic categorisation of faces
Rojahn & Pettigrew (1992)	Is there a memory advantage for schema-relevant information?
Schaubroeck & Muralidhar (1991)	Effects of tabular and graphic display formats on decision-making performance
Schwaighofer, Fischer & Bühner (2015)	Transfer of working memory training
Silverman (2003)	Gender differences in delay of gratification
Sommer, Aleman, Bouma & Kahn (2004)	Gender differences in bilateral language representation
Stadler, Becker, Gödker, Leutner & Greiff (2015)	Relationship between complex problem solving and intelligence
te Nijenhuis, Jongeneel-Grimen & Kirkegaard (2014)	Are Headstart gains on the <i>g</i> factor?
te Nijenhuis, van den Hoek & Armstrong (2015)	Spearman's hypothesis and Amerindians – differences between groups on subtests of IQ battery are a function of the <i>g</i> loadings of subtests
Tybur, Laakasuo, Ruff & Klauke (2016)	How pathogen cues shape impressions of foods
Voss, Kramer, Basak, Prakash & Roberts (2010)	Do expert athletes show cognitive expertise?
Voyer (2011)	Time limits and gender differences in mental rotation
Voyer, Jansen (2017)	Influence of motor expertise on performance in spatial tasks
Voyer, Postma, Brake & Imperato-McGinley (2007)	Gender differences in object location memory
Wang, Liu, Zhu, Meng, Li & Zuo (2016)	Effect of action video game training on cognitive ability of healthy adults

## Organisational Psychology

---

Banks, Batchelor, Seers, O'Boyle, Pollack & Gower (2014)	Relative importance of team-member and leader-member social exchange
Banks, McCauley, Gardner & Guler (2016)	Comparing authentic and transformational leadership
Beal, Cohen, Burke & McLendon (2003)	Relationship between cohesion and performance in groups
Berry, Sackett & Landers (2007)	Relationship between interviews and cognitive ability
Beus & Whitman (2012)	Relationship between typical and maximum performance
Boer, Deinert, Homan & Voelpel (2016)	Role of leader-member exchange in transformational leadership
Bowling, Khazon, Meyer & Burrus (2015)	Situational strength as a moderator of job satisfaction and performance
Burke, Salvador, Smith-Crowe, Chan-Serafin, Smith & Sonesh (2011)	Influence of hazards and safety training on learning and performance
Butts, Casper & Yang (2013)	Importance of work-family support policies for employee outcomes
Choi, Oh & Colbert (2015)	Organisational commitment – roles of personality and culture
Cohen & Hudecek (1993)	Relationship between organisational commitment and turnover
Costanza, Badger, Fraser, Severt & Gade (2012)	Generational differences in work-related attitudes
de Wit, Greer & Jehn (2012)	Relationship between intragroup conflict and group outcomes
Donovan & Radosevich (1998)	Role of goal commitment on goal difficulty-performance relationship
Driskell, Willis & Copper (1992)	Effect of overlearning on retention
Ernst Kossek, E. & Ozeki, C. (1998)	Relationships between work-family conflict, policies, and job-life satisfaction
Foels, Driskell, Mullen & Salas (2000)	Effects of democratic leadership on group member satisfaction and integration
Gonzalez-Mulé, Mount & Oh (2014)	Relationship between general mental ability and non-task performance

Grijalva, Newman (2015)	Narcissism as a predictor of counterproductive work behaviour
Harari, Reaves & Viswesvaran (2016)	Relationships of creative and innovative performance with task, citizenship and counterproductive job performance
Harari & Rudolph (2017)	Effect of rater accountability on performance ratings
Harms, Credé, Tynan, Leon & Jeung (2016)	Relationship between leadership and stress
Jackson & Schuler (1985)	Correlates of role ambiguity and role conflict in work settings
Jiang, Liu, McKay, Lee & Mitchell (2012)	Job embeddedness as a predictor of turnover
Jin & Rounds (2012)	Relationship between stability and change in work values
Joseph, Dhanani, Shen, McHugh & McCord (2015)	Relationship between leader trait affectivity and multiple leadership criteria
Keim, Landis, Pierce & Earnest (2014)	Predictors of job insecurity
Klein, DiazGranados, Salas, Le, Burke, Lyons & Goodwin (2009)	Impact of team-building components (goal setting, interpersonal relations, problem solving, role clarification) on cognitive, affective, process and performance outcomes
Knight & Eisenkraft (2015)	Effects of group affect on social integration and task performance
Koch, D'Mello & Sackett (2015)	Gender stereotypes and bias in employment decision making
Koh, Shen & Lee (2016)	Black-white mean differences in job satisfaction
Liu, Jiang, Shalley, Keem & Zhou (2016)	Motivational mechanisms of employee creativity
Liu, Guo & Chi (2015)	Antecedents and performance consequences of proactive environmental strategy
Martocchio & O'Leary (1989)	Sex differences in occupational stress
Murphy & Athanasou (1999)	Effect of unemployment on mental health

Nicolaides, LaPort, Chen, Tomassetti, Weis, Zaccaro & Cortina (2014)	Relationship between shared leadership and team performance
Orlitzky & Hirokawa (2001)	Predictors of group decision-making effectiveness
Park, Jacob, Wagner & Baiden (2014)	Job control and burnout – testing Conservation of Resources model
Parks & Steelman (2008)	Effectiveness of organisational wellness programs on absenteeism and job satisfaction
Rauch, Rosenbusch, Unger & Frese (2016)	Effectiveness of cohesive and diversified networks
Robertson (1989)	Work-sample tests of trainability predict short-term training success
Rockstuhl, Dulebohn, Ang & Shore (2012)	Correlates of leader-member exchange
Sadri & Robertson (1993)	Relationship between self-efficacy and work-related behaviour
Schmidt, Roesler, Kusserow & Rau (2014)	Link between role ambiguity and role conflict, and depression
Theeboom, Beersma & van Vianen (2013)	Effects of coaching on individual level outcomes in an organisational context
Thorsteinson (2003)	Job attitudes of part-time versus full-time workers
Vaden & Hall (2005)	Effect of simulator platform motion on pilot training transfer
Verquer, Beehr & Wagner (2003)	Relationship between person-organisation fit and work attitudes
Whelpley & McDaniel (2016)	Relationship between self-esteem and counterproductive work behaviours
Wiersma (1992)	Effects of extrinsic rewards in intrinsic motivation

### **Social Psychology**

Anderson & Leaper (1998)	Gender effects on conversational interruption
Anestis, Soberay, Gutierrez, Hernández & Joiner (2014)	Link between impulsivity and suicidal behaviour

Banas & Rains (2010)	Effectiveness of inoculation theory in conferring resistance to persuasion and attitude change
Batalha, Reynolds & Newbigin (2011)	Gender differences in social dominance orientation
Blanken, van de Ven & Zeelenberg (2015)	Assessing the magnitude and moderators of the moral licensing effect
Blondé & Girandola (2016)	Effect of vividness on persuasion
Bornstein & Cecero (2000)	Relationship between interpersonal dependency and five factor model domain scores
Braithwaite & Corr (2016)	Effect of self-efficacy and self-confidence on academic capabilities in university students
Byron & Khazanchi (2011)	Relationship of state and trait anxiety to performance on figural and verbal creative tasks
Calin-Jageman & Caldwell (2014)	Effect of superstition on performance
Costello & Hodson (2014)	Lay beliefs about causes of and solutions to dehumanisation and prejudice – the role of human-animal relations
Craddock, vanDellen, Novak & Ranby (2015)	Relationship between social control and health outcomes
DeCoster & Claypool (2004)	Priming effects on impression formation
Deshpande & Viswesvaran (1992)	Effectiveness of cross-cultural training of expatriate managers
Devine & Philips (2001)	Relationship between cognitive ability and team performance
Donahue (1985)	Relationship between intrinsic and extrinsic religiousness
Drescher & Schultheiss (2016)	Gender differences in implicit affiliation and intimacy motivation
Edens, Campbell & Weir (2007)	Relationship between youth psychopathy and criminal recidivism
Georgesesen & Harris (1998)	Effects of power on self and other performance evaluations



Gerber & Wheeler (2009)	Being rejected frustrates psychological needs
Goemans, van Geel & Vedder (2015)	Longitudinal developmental outcomes of foster children
Gully, Devine & Whitney (2012)	Cohesion and performance – effects on level of analysis and task interdependence
Hauch, Sporer, Michael & Meissner (2016)	Effect of training on detection of deception
Hu, Zhang & Wang (2015)	Relationship between trait resilience and mental health
IJzerman, Blanken, Brandt, Oerlemans, Van den Hoogenhof, Franken & Oerlemans (2014)	Sex differences in distress from infidelity in early adulthood and later life
Leach & Cidam (2015)	Link between shame and constructive approach orientation
Lee, Moon & Feeley (2016)	Effectiveness of legitimisation of paltry favours compliance strategy on compliance rate, mean contribution amount, and total contribution amount
Littleton, Bye, Buck & Amacker (2010)	Relationship between psychosocial stress during pregnancy and perinatal outcomes
Masi, Chen, Hawkley & Cacioppo (2010)	Effectiveness of interventions to reduce loneliness
McMahan & Estes (2015)	Effect of contact with natural environments on positive and negative affect
Morgan, Flora, Kroner, Mills, Varghese & Steffan (2012)	Effectiveness of interventions for treating offenders with mental illness
Mullen & Hu (1988)	Social projection as a function of cognitive mechanisms
Nadler & Clark (2011)	Effects of stereotype threat nullification among African Americans and Hispanic Americans
O'Connor, Morrison, McLeod & Anderson (1996)	Relationship between gender and belief in a just world
Payne & Marcus (2008)	Efficacy of group psychotherapy for older adults

Randall & Wolff (1994)	The time interval in the intention-behaviour relationship
Rise, Sheeran & Hukkelberg (2010)	The role of self-identity in the Theory of Planned Behaviour
Rivis, Sheeran & Armitage (2009)	Relationship of anticipated affect and moral norms to intentions in the Theory of Planned Behaviour
Roccato & Ricolfi (2005)	Relationship between right-wing authoritarianism and social dominance orientation
Roisman, Holland, Fortuna, Fraley, Clausell & Clarke (2007)	Relationship between Adult Attachment Interview and self-reports of attachment style
Schütz & Six (1996)	Relationship between prejudice and discrimination
Sibley, Osborne & Duckitt (2012)	Personality and political orientation – testing Threat-Constraint model
Stebly, Besirevic, Fulero & Jimenez-Lorente (1999)	Effects of pre-trial publicity on juror verdicts
Terrizzi, Shook & McDaniel (2013)	Behavioural immune system as a predictor of social conservatism
van Osch, Blanken, Meijs & van Wolferen (2015)	Evaluating evidence for the Group Attractiveness Effect
Vazire & Funder (2006)	Relationship between impulsivity and the self-defeating behaviour of narcissists
Vedel (2014)	Relationships between the Big Five and tertiary academic performance
Vicaria & Dickens (2016)	Intra- and Inter-personal outcomes of interpersonal coordination
Zuckerman, Li & Hall (2016)	Gender differences in self-esteem
<b>Close Replications</b>	
Cheung, Campbell, LeBel, Ackerman, Aykutoğlu, Bahník, ... Caprariello (2016) – RRR5	Exposure to a commitment prime affects subjective commitment to a relationship

---

	Commitment prime and exit from a relationship
	Commitment prime and perceived neglect in a relationship
	Commitment prime and communicating in a relationship
Ebersole, Atherton, Belanger, Skulborstad, Allen, Banks... Boucher (2016) – Many Labs 3	Commitment prime and relationship loyalty Availability heuristics –differences in estimating frequency of easier versus harder to imagine words
	Relationship between persistence and conscientiousness
	Metaphoric restructuring – priming can influence interpretation of an ambiguous temporal statement
	Credentials and prejudice – are prejudicial attitudes more likely when prior behaviour shows people were non-prejudiced?
	Can power impair perspective taking?
	Differences in self-esteem and subjective distance of events
	Impact of argument strength on persuasion (elaboration likelihood)
	Stroop effect – difference between congruent versus incongruent conditions

	Does physical warmth affect perceptions of personality?
	Weight embodiment – does holding heavier objects influence perceptions of importance
Eerland, Sherrill, Magliano, Zwaan, Arnal, Aucoin... Carlucci (2016) – RRR3	Grammatical aspect (use of imperfective aspect versus perfective aspect) can affect perceptions of criminal intentionality
	Grammatical aspect affects detailed processing
	Grammatical aspect affects intention attribution
Hagger, Chatzisarantis, Alberts, Anggono, Batailler, Birt... Bruyneel (2016) – RRR4	Ego depletion affects reaction time
	Ego depletion affects reaction time variability
Klein, Ratliff, Vianello, Adams, Bahnik, Bernstein... Brumbaugh (2014) – Many Labs 1	Sunk costs – difference between likelihood to attend an event if financial investment has been made versus if event is free
	Gambler's fallacy – does witnessing a rare chance event influence beliefs about prior events
	Flag priming – exposure to a flag may increase conservatism
	Quote attribution – source of information affects how information is perceived and evaluated

Currency priming – exposure to money  
increases endorsement of the social system

Imagined contact – imagining contact with  
outgroups can reduce prejudice

The effect of gain versus loss framing on  
willingness to take risks

Allowed/forbidden – question phrasing can  
influence responses

Norm of reciprocity – decisions about  
allowing/denying behaviour of an outgroup  
are affected by responses regarding  
behaviour of ingroup

Low versus high category scales – different  
response options influence self-assessment  
of behaviour

Sex differences in math attitudes

Relation between implicit and explicit math  
attitudes

Anchoring – estimating a number differs  
between large and small anchors (1 –  
distance to New York City; 2 – population of  
Chicago; 3 – height of Mount Everest; 4 –  
Babies born per day)

Klein, Vianello, Hasselman, Adams, Adams,  
Alper... Bahník (2018) – Many Labs 2

Associations between cardinal direction and  
socioeconomic status

Structure promotes goal pursuit

Disfluency engages analytic processing

Consumerism undermines trust

Influence of incidental anchors on judgment

Associations between sociometric status and wellbeing

False consensus: supermarket scenario – false consensus regarding how common own responses are among other people

False consensus: traffic-ticket scenario

Associations between vertical position and power

Reluctance to tempt fate –testing the belief that tempting fate increases bad outcomes

Differences in preferences for formal versus intuitive reasoning

Less-is-better effect – gifts are compared in price relative to similar gifts

Moral typecasting – attribution of intentionality and responsibility relating to perceived moral agency

Can moral violations induce desire for cleansing?

	Does priming “heat” increase belief in global warming?
	Are helpful and harmful side effects differentially perceived as being intended?
	Relationship between directionality and similarity
Wagenmakers, Beek, Dijkhoff, Gronau, Acosta, Adams... Blouin-Hudon (2016) – RRR6	Facial feedback hypothesis – people’s affective responses can be influenced by their own facial expression

---

### 3.2.2 Data extraction and analysis

Where the results of more than one meta-analysis were reported, the one including the largest number of studies was extracted. If multiple meta-analyses included the same number of studies, the first was used.

Heterogeneity for each meta-analysis was computed using the DerSimonian-Laird estimator (see section 2.1.5 for details) in Metafor in R (Viechtbauer, 2010)<sup>5</sup>. In order to keep effect sizes and levels of heterogeneity consistent across studies, all effect sizes were input as Cohen’s  $d$ . All other effect sizes were converted accordingly.

Once mean  $d$  and  $T$  are determined for a given meta-analysis, the flop index ( $FI$ ) can be computed. Our computation is based on the widely held assumption that the distribution of true effect sizes is normal (Hunter & Schmidt, 2004). In this case,  $FI$  equals the proportion of scores  $< 0$  in a normal distribution, with mean  $d$  and standard deviation  $T$  (see Figure 3.2, shaded area).

---

<sup>5</sup> We also computed analyses using the widely used Hunter-Schmidt and the Paule-Mandel estimators. These led to similar results (see Appendix A) and the same conclusions.

For a random sample of 15 meta-analyses, relevant data were independently extracted and heterogeneity and  $FI$  recalculated by the principal supervisor to establish reliability. This resulted in correlations of  $r = .97$  for  $T$ ,  $r = 1.00$  for  $d$ , and  $r = 1.00$  for  $FI$ .

It turned out that the frequency distributions for some of our observed outcome variables were right skewed. For example, among the 150 meta-analyses,  $T$  had a skewness 0.99 (the largest  $Z$  score being 3.57). We therefore report winsorized means ( $M_{win}$ ) and standard deviations ( $SD_{win}$ ).  $M_{win}$  removes the undue effect of outliers, but retains much more information than the median, which trims all scores but the one in the middle of the distribution (Erceg-Hurn & Mirosevich, 2008). Specifically for  $T$ ,  $M_{win}$  should also counteract the likely overestimation resulting from setting negative heterogeneity estimates to  $T = 0$ . We use the Yuen-Welch method (Wilcox, 2005), which is similar to the t-test but based on  $M_{win}$ , for group comparisons of  $T$ . Similarly, we use winsorized correlations,  $r_{win}$  (Wilcox, 2005). These limit the effect of outliers, but retain more information than Spearman's rank-based correlation,  $r_s$ . All data and further materials can be found at <https://osf.io/yr3xd/>.

### 3.3 Results

#### 3.3.1 How meta-analyses address heterogeneity

Out of 150 meta-analyses, 123 tested moderators, but only 83 (55%) reported a measure of heterogeneity. In 2009, the influential PRISMA guidelines (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, Moher, Liberati, Tetzlaff, & Altman, 2009) recommended that meta-analyses should address heterogeneity. But even in post-2009 meta-analyses, heterogeneity was only reported in 60% of cases. Interestingly, statistical significance of heterogeneity, e.g.  $Q$ , was widely reported (77 times); of those, 45% did not report quantifying information. This focus on statistical significance and neglect of quantifying information runs counter to the meta-analysis estimation perspective (Hunter, 1997). Overall, heterogeneity was quantified in less than a third of cases (43 times out of 150):  $I^2$  was reported in 33 cases,  $T^2$  in 9, and another measure was reported once. In addition to the observed neglect of quantification, it is interesting that authors unanimously reported  $T^2$  (the heterogeneity variance) instead of  $T$  (the standard deviation). Whereas standard deviation has a meaning that is comparatively easy to grasp (it approximates the average difference from the mean), variance does not have a similarly accessible interpretation. (This is why researchers most commonly report standard deviations, and not variances, in their descriptive statistics. Similarly, we can picture standard deviations in Figure 3.2, but would be lost with variances.)



This might suggest that even authors who quantify heterogeneity often do not fully recognize the value of reporting  $T$ .

### 3.3.2 Heterogeneity observed in close replications and conceptual replications

Table 3.2 shows descriptive statistics for close replications and conceptual replications. As expected, heterogeneity was much lower ( $M_{\text{win}} = 0.09$ ) in close replications than in the conceptual replications ( $M_{\text{win}} = 0.33$ ),  $t(94.9) = 10.43$ ,  $p < .001$  (see Figure 3.4). Whereas Stanley et al.'s (2018) estimate of heterogeneity in conceptual replications (median  $T = 0.35$ ) is virtually the same as ours, the same is not true for close replications. Their much higher estimate ( $T = 0.25$ ), was based on only two Many-Lab type replications (Eerland et al., 2016; Hagger et al., 2016), both of which were part of our sample.

Table 3.2 Descriptive Statistics for Study 1a.

Sub-discipline	No. meta-analyses	Mean $k$ per meta-analysis	Mean $T$ ( $SD$ )	Mean $FI$ ( $SD$ )
Close replications	57	35.2	0.09 (0.07)	0.01 (0.01)
Social	50	35.7	0.31 (0.11)	0.06 (0.06)
Cognitive	50	36.5	0.32 (0.13)	0.07 (0.08)
Organisational	50	38.3	0.35 (0.10)	0.08 (0.07)
<i>Total</i>	<i>150</i>	<i>36.9</i>	<i>0.33 (0.11)</i>	<i>0.08 (0.07)</i>

*Note.* All means are winsorized. Means for Flop Index are only for meta-analyses which are statistically significant ( $p < .05$ ).

Contrary to our hypothesis, levels of heterogeneity were very similar across all three sub-disciplines (cognitive versus social psychology:  $t(57.3) = 0.33$ ,  $p = .370$  (1-tailed); social versus organisational:  $t(57.0) = 1.17$ ,  $p = .125$  (1-tailed); organisational versus cognitive:  $t(54.9) = 0.74$ ,  $p = .463$  (2-tailed).

What do average  $T$ s of 0.09 (close replications) and 0.33 (conceptual replications) mean? As we discussed earlier, this partly depends on the effect size, which averaged  $d = 0.31$  and  $d = 0.47$  (winsorized means for close replications and conceptual replications<sup>6</sup>, respectively). The average result for conceptual replications is depicted in Figure 3.2 (solid line). Remember that

<sup>6</sup> This is again very close to Stanley et al.'s (2018) average effect size of  $d = 0.39$ .

Cohen's  $d$  of 0.2/0.5/0.8 are often considered as benchmarks for small/medium/large effects. All of these occur frequently in the distribution of true effect sizes. Moreover,  $FI = 0.09$  reflects that 9% of true effect sizes even run against expectation. We will discuss the meaning of these results in more detail later, but at first glance observed heterogeneity in conceptual replications appears large. In contrast, the average close replication showed little variability in population effect sizes and  $FI$  was virtually zero.

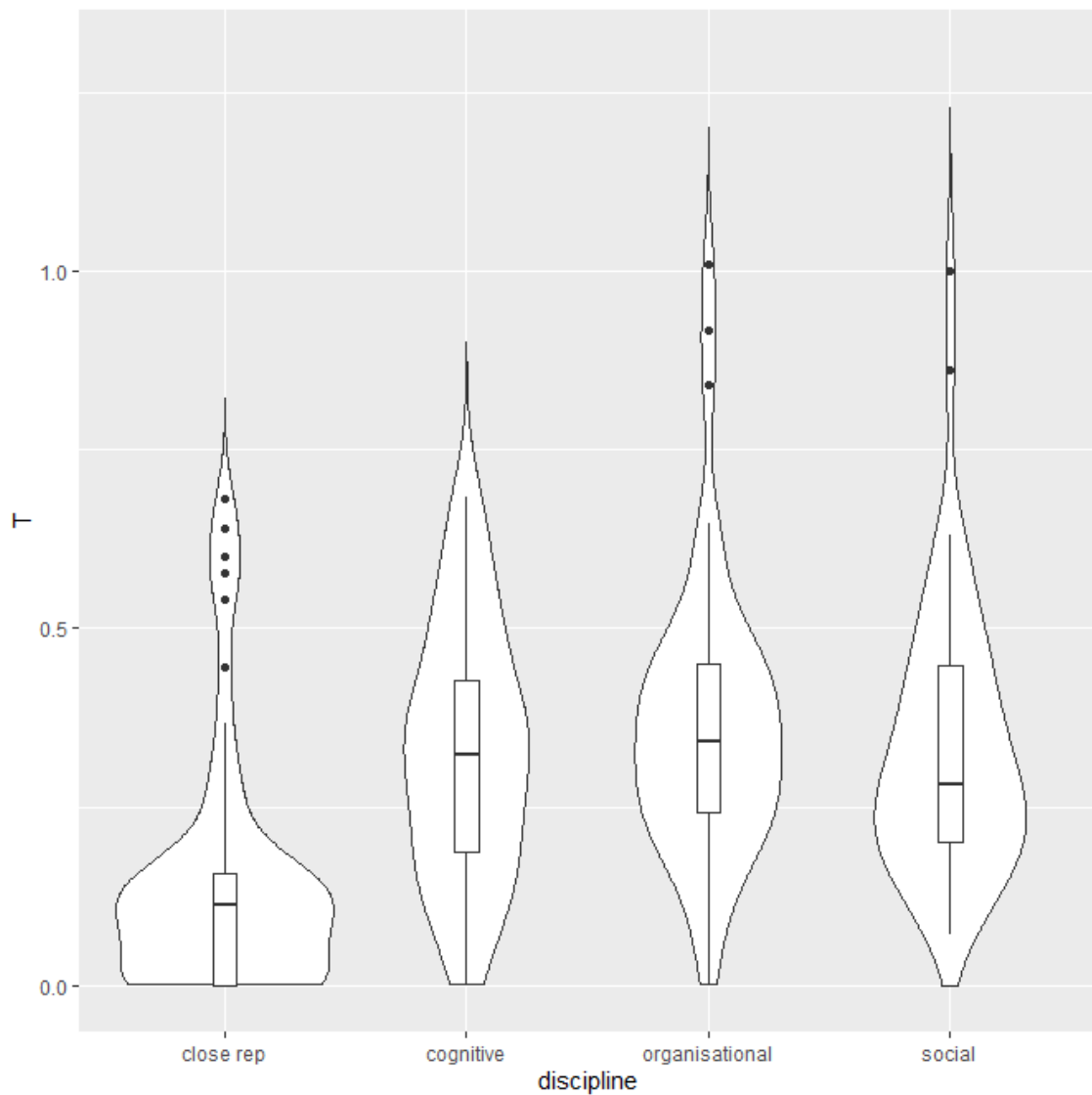


Figure 3.4 Observed levels of heterogeneity for 57 close replications and for 50 meta-analyses (each) in cognitive, organisational and social psychology, with box-plots for  $M_{win}$ .

### 3.3.3 Moderators

In order to investigate to what extent moderators account for heterogeneity, we looked at all 36 meta-analyses with  $k \geq 60$ , because moderators will be most reliably identified in large

meta-analyses. Where possible, we used the strongest moderator for which sufficient information was reported for further analyses. All moderators thus identified were grouping variables (i.e. none was continuous). We used these moderators to partition studies into appropriate subsets, which left us with 22 meta-analyses. We then excluded broad subsets. For example, Baker, Peterson, Pulos, and Kirkland, R. A (2014) looked at intelligence and “Reading the Mind in the Eyes”. The strongest moderator they examined was the type of intelligence test used. Based on this, studies were split into two subsets: IQ measured using the Wechsler IQ test, and IQ measured using any other test. We excluded the broad ‘other’ subset and compared  $T$  from the subset of studies measured using the Wechsler IQ test against the  $T$  in the initial overall meta-analysis. Average  $T$  in the 22 subsets ( $M_{\text{win}} = 0.33$ ) was very similar to average  $T$  in the corresponding 22 overall meta-analyses ( $M_{\text{win}} = 0.37$ ),  $t(13) = 1.39$ ,  $p = .187$ . Powerful moderators might only emerge when they are based on theoretical considerations. We therefore looked at those 10 out of the 22 meta-analyses that presented a theoretical rationale for the moderator. Again, we used these moderators to split the studies from each meta-analysis into appropriate subsets and repeated the previous analysis. We found that average  $T$  in the 10 moderator-based subsets ( $M_{\text{win}} = 0.37$ ) was again very similar to average  $T$  in the 10 corresponding overall meta-analyses ( $M_{\text{win}} = 0.37$ ;  $t(5) = 0.73$ ,  $p = .499$ ).

This moderator analysis does not suggest that the large heterogeneity in our conceptual replication sample is readily explained by mixing apples and oranges. Still, the possibility remains that authors (potentially unwisely) combine highly diverse studies and then fail to address relevant moderators. In order to address this point, we rated (on a single, global five-point scale, from ‘low’ to ‘high’) for each conceptual replication how broad or narrow its inclusion criteria were. Ratings considered to what extent the addressed question was broad (e.g. ‘How effective is psychotherapy?’) versus narrow (e.g. ‘How effective is cognitive behavioural therapy to treat simple phobias?’); to what extent the manipulation of the IV and the measurement of the DV followed a standard protocol; and the similarity of the samples included. Ratings were conducted by the principal supervisor without knowledge of the actual levels of heterogeneity; to establish reliability, a random sample of 30 meta-analyses were independently re-rated by the principal investigator. We computed inter-rater agreement as Cohen’s kappa using quadratic weights and observed  $\kappa_w = 0.73$ , which is typically interpreted as good (Jakobsson & Westergren, 2005). For the 58 meta-analyses whose broadness of inclusion criteria was rated ‘low’ or ‘low-to-medium’, average heterogeneity was still very high ( $M_{\text{win}} = 0.29$ ). In other words, if authors generally avoided meta-analyses that integrate fairly diverse studies, levels of observed heterogeneity would probably not be much lower. In sum,

our analyses do not support the view that unwise mixing of apples and oranges is a strong driver of observed heterogeneity in conceptual replications.

### 3.3.4 Exploratory analyses on what drives heterogeneity

Heterogeneity differed substantially between conceptual replications ( $SD_{win} = 0.11$ ). This begs the question, why? A number of ideas have been proposed that we can test here. Kenny and Judd (2019) suggested that research areas with larger average effect sizes should have greater levels of heterogeneity. We therefore correlated mean  $d$  with  $T$ . In support of this idea, we found a strong correlation  $r_{win} = .49$  ( $p < .001$ ) for the set of 150 conceptual replications (see Figure 3.5). This replicated across all three sub-disciplines (cognitive:  $r_{win} = .34$ ,  $p = .019$ ; social:  $r_{win} = .52$ ,  $p < .001$ ; organisational:  $r_{win} = .62$ ,  $p < .001$ ). Nonetheless,  $FI$  showed a strong negative relationship with  $d$ ,  $r_{win} = -.56$  ( $p < .001$ ) as one would expect. The relationship between mean  $d$  with  $T$  also held for the set of 57 close replications,  $r_{win} = .48$  ( $p < .001$ ), see Figure 3.5. In light of the link between mean  $d$  and  $T$ , the direct comparison of heterogeneity in close replications versus conceptual replications might appear doubtful. This is because mean  $d$  proved considerably lower in close replications (0.31) than in conceptual replications (0.47). Correction via the relevant regression equation (Figure 3.5) suggests an average  $T$  of 0.27 for conceptual replications at  $d = 0.31$  (the average for close replications). This is still much larger than that observed for close replications ( $M_{win} = 0.09$ ); moreover, at  $d = 0.31$ ,  $FI$  rises to 13% for the average conceptual replication (compared to  $FI = 0\%$  for the average close replication).

Looking at systematic reviews in healthcare, IntHout et al. (2015) found that small studies are more heterogeneous (measured as  $T^2$ ) than large ones. We therefore worked out the median sample size for each conceptual replication and correlated this with  $T$ . This resulted in  $r_{win} = -.10$  ( $p = .108$ , 1-tailed), not suggesting an important role for average sample size.

Richard et al. (2003) proposed that as a research field matures, the focus shifts from establishing an effect to exploring its boundaries, and this should increase heterogeneity in findings. If we accept the number of studies ( $k$ ) as a proxy for the maturity of a research field, we should expect a positive correlation between  $k$  and  $T$ . In line with this idea, they found a correlation of  $r = .11$  in a large survey of meta-analyses in social psychology. For our 150 conceptual replications, we found  $r_{win} = .23$  ( $p = .005$ ) in support of this idea. An obvious alternative interpretation is that meta-analyses with broader inclusion criteria cast a wider net and will therefore include more studies than those that use narrow inclusion criteria (Murphy, 2017).

We thought to test these two competing explanations (exploring boundaries versus broader inclusion criteria). If later research into an effect tends to explore its boundaries, we would expect to see higher heterogeneity in studies conducted late than in those conducted early. We therefore looked at all meta-analyses that capture a sufficiently mature research area. We operationalised this by including meta-analyses with  $k \geq 30$  and a time range for included studies of  $\geq 10$  years, which led to the inclusion of 82 meta-analyses. For each of these, we then determined  $T$  separately for the earlier and the latter half of the included studies. Although the difference was in the expected direction (early:  $M_{\text{win}} = 0.336$ ,  $SD_{\text{win}} = 0.097$ ; late:  $M_{\text{win}} = 0.342$ ,  $SD_{\text{win}} = 0.110$ ), it was small and not statistically significant ( $t(49) = 1.01$ ,  $p = .319$ ), therefore not supporting the idea of boundary exploration.

If the observed correlation between  $k$  and  $T$  is down to broader inclusion criteria, we would expect to see that meta-analyses with broader inclusion criteria show higher  $T$ . We therefore correlated our ratings of the broadness of inclusion criteria with  $T$  and found  $r_{\text{win}} = .12$  ( $p = .142$ ). This does not offer strong evidence that the observed correlation between  $k$  and  $T$  reflects broadness of inclusion criteria. In sum, our data do not offer a clear explanation why  $T$  tends to increase with  $k$ .

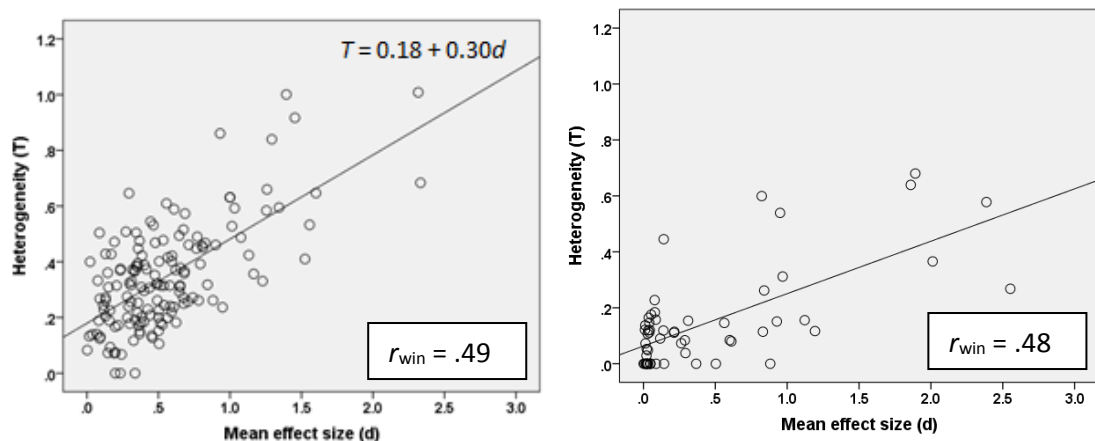


Figure 3.5 Heterogeneity as a function of meta-analyses' mean effect size. Left hand panel: 150 meta-analyses from cognitive, social, and organisational psychology; right hand panel: 57 meta-analyses for close replications.

### 3.4 Discussion

Heterogeneity in research results has received increased attention in the last few years (Kenny & Judd, 2019; McShane & Böckenholt, 2014; Shrout & Rodgers, 2018; Stanley et al., 2018; Van Erp et al., 2017), partly motivated by the discussion of replicability of research findings in psychology (Zwaan et al., 2018). Here we compared heterogeneity between close and conceptual replications and across three psychological subfields. Using Cohen's  $d$  as an effect-size measure, we found a mean  $T$  of 0.33 ( $SD = 0.11$ ) for 150 meta-analyses, which can be typically considered to be collections of conceptual replications (Schmidt & Oh, 2016). In contrast to conceptual replications, close replications typically showed strikingly low heterogeneity (mean  $T = 0.09$ ,  $SD = 0.07$ ).

#### 3.4.1 Heterogeneity and the informativeness of meta-analyses

What does heterogeneity tell us about the result of an individual meta-analysis? We use two examples from cognitive psychology for an illustration. With the importance of working memory for second language proficiency development and processing being debated, Linck, Osthus, Koeth, and Bunting (2014) investigated the strength of this link in a meta-analysis. Included studies used a range of working memory tasks and second language comprehension measures in diverse samples. The strength of the relationship proved medium in size ( $d = 0.51$ ), and heterogeneity was estimated to be low ( $T = 0.11$ ) (see Figure 3.1, left panel). The latter implies high consistency of the relationship and ready generalisability across paradigms. In line with this, most true effect sizes (mean  $\pm 1SD$ ) should fall in a narrow range of medium sized effects ( $d = 0.40$  to  $0.62$ ).

Baker, Peterson, Pulos, and Kirkland (2014) used meta-analysis to investigate the degree of independence between general intelligence and mental state understanding. Included studies used a range of established intelligence tests in diverse samples, however, all used the same widely used test of mental state understanding (Reading the Mind in the Eyes Test). As in the previous example, the strength of the relationship proved medium in size ( $d = 0.49$ ), however, heterogeneity was estimated to be much higher ( $T = 0.35$ ) (Figure 3.1, right panel), even though the same test of mental state understanding was used throughout. Even though the observed level of heterogeneity was medium and not large, it already implies low consistency of the studied relationship and a lack of generalisability across paradigms. Most true effect sizes (mean  $\pm 1SD$ ) should fall in a wide range of very small to large effects ( $d = 0.14$  to  $0.84$ ). The authors reported a statistically significant moderator, but given its a-theoretical nature,

and the number of moderators tested, it remains debatable if this reflects progress in understanding, or successful capitalisation on chance (Ioannidis, 2008).

In sum, it appears that the relationship between working memory and second language proficiency is better understood than that between intelligence and performance on the Reading the Mind in the Eyes Test. More generally, everything else being equal, meta-analyses with lower heterogeneity will be more informative. Erroneous effect sizes will inflate heterogeneity in a meta-analysis. Incidentally, this will reward meta-analysts' accuracy once heterogeneity is considered more routinely.

Richard et al. (2003) suggested that observed heterogeneity is not necessarily informative in itself. Where studies in a research area intend to verify a particular effect, it will reflect consistency; but where a lot of studies seek to reverse the effect, it will reflect manipulability. However, in our two examples from cognitive psychology above, no such difference was at play. Moreover, where the failure or even reversal of an effect is properly understood (instead of occasionally observed), a powerful moderator should emerge that allows to split studies into fairly homogenous subsets. Our moderator analysis suggests that this is rarely the case.

### 3.4.2 Replicability

Our results have important implications for the evaluation of observed replication rates in psychology (see also Stanley et al., 2018). Open Science Collaboration (2015) famously attempted close replications of 100 recent studies. Although using larger samples than the original studies, statistical significance was achieved in only 36% of replications (25% in social psychology and 50% in cognitive psychology). This finding has become a catalyst of the controversial debate about the health of psychology research, which is still ongoing (e.g. Earp & Trafimow, 2015; Pashler & Harris, 2012; Schmidt & Oh, 2016; Simons, 2014; Stroebe & Strack, 2014). This is not the place to review this debate in full (see Zwaan et al., 2018, for a comprehensive summary). Briefly, a particularly critical reading of these findings suggests that effect size inflation and type 2 errors are rife in the published literature. Two arguments made in defence of the original studies concern hidden moderators and statistical power.

Even close replication attempts must, necessarily, involve some differences to the original study. For example, the perceived funniness of a cartoon may differ across different groups of participants. Therefore failure to replicate could be attributed to these differences (hidden

moderators)<sup>7</sup>. However, as our result showed, heterogeneity tends to be small (mean  $T = 0.09$ ) for multiple close replications. This demonstrates that producing reliable results is possible in practice. If one accepts the status of Open Science Collaboration (2015) replications as close replications, then hidden moderators cannot account for many of the observed replication failures.

A similar argument to defend the validity of studies that failed to replicate draws on statistical power (Stanley et al., 2018). The latter is reduced by heterogeneity (Kenny & Judd, 2019; McShane & Böckenholt, 2014). Unavoidable differences between original studies and close replications might introduce heterogeneity, and thereby reduce the statistical power of the replication studies to a critical extent. Based on only two Many-labs type close replications, Stanley et al. estimated heterogeneity to be high, and subsequent power calculations demonstrated that heterogeneity should therefore decrease power in the typical psychological study to levels that are in line with the low replication rate observed in Open Science Collaboration's (2015). However, heterogeneity of the order we observed in a much larger sample for close replications (mean  $T = 0.09$ ) reduces statistical power only marginally. If we stick to sample sizes that generate 80% power at zero heterogeneity, power does not drop at all for large effects, drops to 78% for medium effects, and to 71% for small effects, as McShane and Böckenholt (2014) showed. The mean effect size for the original studies included in the Open Science Collaboration (2015) was large ( $d = 0.87$ )<sup>8</sup>. Therefore replication power should not be greatly affected, provided that the differences between the Open Science Collaboration replication studies and their original counterparts is comparable to the differences in multiple close replications.

Moreover, if replication failure reflects heterogeneity-driven low power, as Stanley et al. (2018) claimed, higher replication failure in social psychology (Open Science Collaboration, 2015) should be reflected in larger heterogeneity in this field. Our findings (virtually identical heterogeneity levels across cognitive, organisational, and social psychology) do not support this view. In conjunction with the low heterogeneity observed in close replications, it strengthens the interpretation that the low replication rate demonstrated in Open Science Collaboration raises questions regarding the prevalence of publication bias and QRPs. In our

---

<sup>7</sup> For a single study, replication failure could obviously represent rotten luck. However, for a larger number of failed replications, this is implausible.

<sup>8</sup> The mean effect size of the original studies underlying the 57 Many-labs type close replications was  $d = 0.75$  ( $SD = 0.37$ ). Regarding the effect size of the underlying original studies, Many-labs type close replications and Open Science Collaboration replications are therefore comparable. This matters because of the observed link between  $T$  and mean effect size (see Figure 3.5).



eyes, this is good news because promising strategies to combat these biases have been developed (Munafò et al., 2017).

### 3.4.3 Research planning

Average levels of heterogeneity ( $T = 0.33$ ) have quite dramatic effects on power: it drops from 80% to 71% for large effects, from 80% to 66% for medium effects, and from 80% to 57% for small effects (Kenny & Judd, 2019). What level of heterogeneity should we expect for a new study that is not a close replication? The researcher's informed judgment will always be necessary, however the following suggestions might appear sensible. Where a relevant meta-analysis reports  $T$ , use this. Where such a meta-analysis only reports the effect size, use  $T = 0.18 + 0.30d$  (see Figure 3.5) to estimate heterogeneity. Where there is no meta-analysis, the heterogeneity can still be estimated (although with lower precision) based on a single study effect size. When we used all effect sizes from all 150 meta-analyses in our sample to predict heterogeneity,  $T = 0.28 + 0.11d$  was the resulting regression ( $R = .38$ ). Finally, when an effect size estimate is not available, use the mean ( $T = 0.33$ ).

### 3.4.4 Generation of novel predictions

Heterogeneity can help to generate novel predictions. We take sex differences and their cross-cultural variability as an example. From an evolutionary psychology perspective, some psychological sex differences evolved as inherited adaptations reflecting the different adaptive problems women and men have faced. For example, males and females might have evolved different mate preferences (Buss, 1989). In contrast, a social structural perspective explains behavioural sex differences by the different roles men and women occupy in society (Eagly & Wood, 1999). Where a sex difference shows low heterogeneity across cultures, this is more easily accommodated by an evolutionary account, whereas high heterogeneity plausibly points to the power of cultural influences. We note that a quantification of cross-cultural heterogeneity is missing in prominent discussions of sex differences (Buss, 1989; Eagly & Wood, 1999; Hyde, 2014; Shackelford, Schmitt, & Buss, 2005). Cross-cultural heterogeneity of physical sex differences might provide a valuable comparison standard (this idea will be explored in greater detail in Study 3).

The observed cross-cultural heterogeneity also puts the magnitude of any sex difference into perspective. We take math performance as an example. Else-Quest, Hyde, and Linn (2010) meta-analysed results from two international large-scale tests of students maths performance:

TIMSS (46 countries) and PISA (40 countries). For TIMSS the sex difference was  $d = -0.01$  (favouring girls), for PISA, the sex difference was  $d = 0.11$  (favouring boys). Do TIMSS and PISA paint a different picture of girls' and boys' math performance? This depends on the observed heterogeneity. If  $T$  was 0.30, they would not appear to differ meaningfully. In contrast, if  $T$  was 0.01, the studies would paint different pictures. (We calculated  $T = 0.13$  and  $T = 0.06$  for TIMSS and PISA, respectively.)

### 3.4.5 Empirical cumulativeness

Science can be described as a quest to explain the apparent complexity of the natural world through simpler, fundamental principles. Empirical cumulativeness reflects the extent to which empirical findings fit such a simple or explicable pattern. *Ceteris paribus*, high levels of (unexplained) heterogeneity indicate lower empirical cumulativeness (Asendorpf et al., 2013; Hedges, 1987; Murphy, 2017; Richard et al., 2003; Sells, 1963). We saw that typical levels of heterogeneity are quite high. To add some perspective, we can compare typical levels of heterogeneity (variability within a specific topic) with the variability in mean effect sizes across meta-analyses (variability between topics). Whereas we found  $T = 0.33$  for the former, for the latter we observed  $SD = 0.42$  across all 150 meta-analyses. In other words, variability within phenomena measured in this way is only about 20% less than variability between phenomena. Remember also that moderators typically did little to account for or explain heterogeneity. The implications for empirical cumulativeness are sobering.

Why is heterogeneity so high in psychology? One reason might be that researchers typically give greater consideration to measurement of the dependent variable than manipulation of the independent variable. Psychometrics is an important field in psychology, and it is difficult to imagine training in psychology that does not address key concepts of measurement such as reliability and validity. The same level of sophistication is not applied to independent variables. Take ego depletion as an arbitrary, but not untypical, example. Ego depletion describes the idea that self-control is a limited resource; once this is spent, performance on tasks dependent on effortful control suffers. In typical tests, only experimental participants engage in an ego-depleting task before addressing the critical task, which also involves self-control and is completed by all participants. Ego-depletion tasks often appear to be created ad-hoc; among others, they require participants to cross out particular letters under particular circumstances, to resist tempting food, or to complete an incongruent Stroop task (Hagger, Wood, Stiff, & Chatzisarantis, 2010). Progress seems least likely where different approaches to the

independent variable can only be described as different. This might be seen as the lowest rung on a ladder of virtue (see Figure 3.6).

Successful application of moderators appears more promising where different approaches to the independent variable, or other critical study characteristics, can be grouped in a theoretically sensible way. For example, Hagger et al. (2010) suggested different types of experimental tasks, achieving ego depletion via the self-control of emotions versus thoughts versus impulses, etc. A more informative model describes moderators on a continuous scale, and not just on a nominal scale as in the ego depletion case. This approach appears rare, especially the use of a theoretically motivated continuous moderator. Else-Quest, Hyde, and Linn (2010) provide a fine example. They showed that variability in sex differences in math performance across countries is partly explained by the degree of gender equity in those countries. The highest rung on our ladder of virtue is achieved where independent variables can be described on a ratio scale. Examples include degrees of genetic similarity, the intensity of physical stimuli, or the probability of events. These proved fundamental for breakthroughs such as the heritability of various psychological traits (e.g. Plomin, 1990), Stevens' Law (which describes the relationship between the physical and perceived intensity of a stimulus; Stevens, 1957); signal detection theory (which describes uncertainty in perception; Tanner Jr & Swets, 1954); and prospect theory (which describes how people make risky decisions; Kahneman & Tversky, 1979) . Obviously, this highest rung might not always be achievable. But we hope it illustrates the benefits that can be gained from a thorough conceptualization of not just dependent variables but also independent variables.

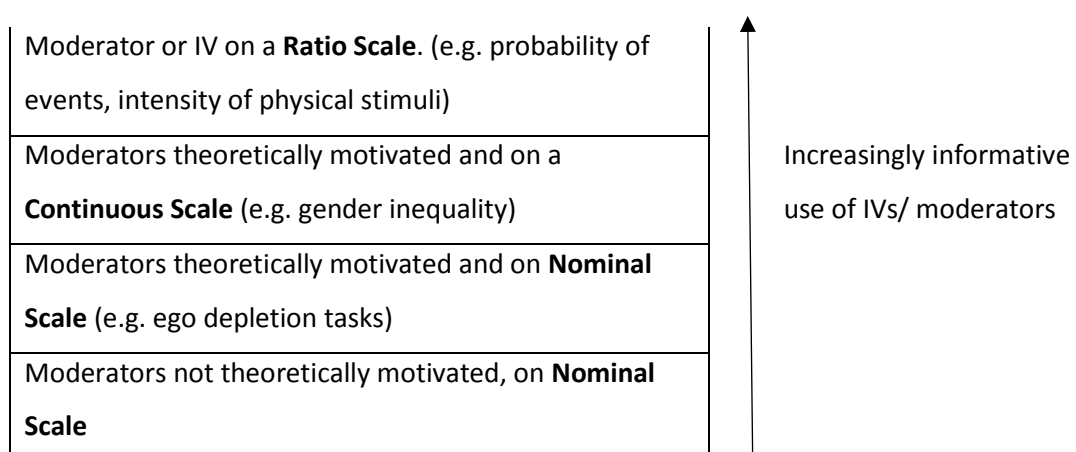


Figure 3.6 Ladder of virtue for independent variables/moderators.

### 3.4.6 Theoretical cumulativeness

#### 3.4.6.1 *Knowledge as a tool*

We can see throughout the history of science how progress in understanding is turned into new tools that drive further theoretical progress. For example, an understanding of electromagnetic fields and sub-atomic particles is essential for the construction of particle accelerators like CERN. The latter then allowed verification of the existence of the Higgs boson, which was predicted by the Standard Model of particle physics. Plausibly, heterogeneity undermines this process (Meehl, 1978). For example, the test of a psychological theory X might require the induction of a particular mood. Where this mood induction shows large heterogeneity, a negative finding of the test can be blamed on (unreliable) methods and theory X is protected from falsification, with detrimental consequences for theoretical progress (Ferguson & Heene, 2012; Greenwald, 2012; Kerr, 1998; LeBel & Peters, 2011; Meehl, 1978).

In this context, the low heterogeneity observed in close replications ( $T = 0.09$ ) is encouraging, as this suggests that closely sticking to an agreed protocol (for both data collection and analysis) typically produces reliable results. Note, however, that these results are based on data from pre-registered studies published irrespective of their results. Consequently, publication bias and QRPs could not have distorted effect sizes. The same is unlikely to hold for published results (Ferguson & Brannick, 2012; Kühberger, Fritz, & Scherndl, 2014; LeBel & Peters, 2011; Levine et al., 2009; Pashler & Harris, 2012; Sterling, 1959), therefore close replications based on published results cannot be expected to be as reliable as this (Open Science Collaboration, 2015). Finally, the results of conceptual replications tend to have low consistency.

In summary, we find the idea of a 'research space' useful (Asendorpf et al., 2013). This is defined by the combination of different manipulations of the independent variable, different dependent variables, different study populations, etc. Where previous findings are used as a research tool, e.g. to induce a particular mood, researchers should be highly selective in their use of the research space, and closely stick to a particular protocol that proved successful. Ideally the suitability of this protocol should have been demonstrated in a pre-registered replication.

#### 3.4.6.2 *Testing theories*

In contrast, testing of theories requires a broad exploration of the research space. Where theories are tested too narrowly, two dangers lurk. First, to evaluate the explicitly stated or implied applicability of a theory, the relevant research space needs to be sampled, e.g. different stimuli sets should be tested, instead of a single standard set. This is the only way to rule out that some theory-irrelevant feature of the standard set drives the observed effect (see Fiedler, 2011, for an example). Second, the explanation offered by a theory might be overly specific (e.g. memory for a word list is improved when the survival value of its items is to be judged). A more general, and thus more parsimonious, explanation should be considered (e.g. memory for a word list is improved by any judgments that trigger self-referent encoding). To test this alternative, it is necessary to test instances of the research space that could violate the overly specific theory while still holding for a more general account (Fiedler, Kutzner, & Krueger, 2012; Shrout & Rodgers, 2018).

A good theory should specify its scope. To evaluate the theory, meta-analytic summary of the relevant studies needs to move beyond focusing solely on the mean effect size and its statistical significance. In addition,  $T$  (or another appropriate quantification of heterogeneity) and *Flop Index* (an estimate of the proportion of true effect sizes that are opposite to the expected direction) are essential. Our review of reporting of heterogeneity in meta-analyses suggests that this is rarely implemented.

#### 3.4.7 Design of applications

One aim of psychological research is to provide a knowledge base for the design of helpful practical applications. Consider enrolment in pension schemes as an example: In order to boost enrolment in voluntary workplace pensions, auto-enrolment (where employees have to actively opt out if they do not want to be a member of the scheme) proved highly successful. For example, enrolment rates in a large US company jumped from 37% to 86% over the course of one year (Madrian & Shea, 2001). Similarly, in the UK, enrolment of private sector employees jumped from 2.8 million to 5.5 million within two years (Office for National Statistics, 2015). The introduction of auto-enrolment was based on status quo bias, the principle that people tend to postpone decisions that they perceive as complex (e.g. Tversky & Shafir, 1992). In the pensions example, the translation of research findings into a practical application was successful. Should we expect that in general? Where heterogeneity is large, the result of the next study (and we might conceptualize a new practical application as the

'next study') is uncertain. But the prediction of how successful a new practical application will be should be easier where heterogeneity is low. Our results suggest two conclusions. First, the ubiquity of high heterogeneity implies that the effectiveness of a new application is difficult to accurately predict. The *Flop Index* of 0.09 for the average meta-analysis suggests that about 9% of its studies even pursue a 'true' effect that is opposite to the expected direction. Second, irrespective of the level of heterogeneity, an application looks more promising when the average effect size is large (compare the two distributions in Figure 3.2).

#### 3.4.8 Limitations

Due to the novelty of the concept and the enormous effort involved, Many-Labs type close replication attempts (Klein et al., 2014) are relatively rare. More importantly, the set of original studies that motivated these replications cannot be considered to be representative of psychological research in general. The observed difference in average effect size (close replications:  $d = 0.31$ , meta-analyses:  $d = 0.47$ ) supports this point. It therefore remains unknown how readily the low heterogeneity observed in close replications would generalise to psychological research findings in general. Also, for feasibility reasons we could only consider meta-analyses for which we could obtain sufficient details regarding the effects in underlying primary studies. It remains unclear to what extent our resulting sample is representative of meta-analyses in the three fields considered and beyond. It is reassuring, though, that previous heterogeneity surveys with markedly different sampling frames observed very similar results (Stanley et al., 2018; Van Erp et al., 2017).

Again for feasibility reasons, we had to rely on reported effect sizes for the underlying primary studies. One systematic investigation of meta-analyses in medicine found that about 1 in 5 effect sizes were erroneous (Gøtzsche, Hróbjartsson, Marić, & Tendal, 2007). This should add (error) variance to the meta-analysis, and consequently inflate observed heterogeneity. In the absence of reliable data for how frequent and large such errors are in meta-analyses in psychology, an appropriate adjustment is currently out of reach.

Finally, our recalculation of each meta-analysis unrealistically treated all studies as between-subjects, and this causes problems for the estimation of heterogeneity. In order to estimate heterogeneity, the observed variability in effect sizes is adjusted by removing the effect of sampling error. Where the design is within-subjects, meta-analytic treatment of the study as between-subjects will overestimate its sampling error. Consequently, heterogeneity calculation over-adjusts, because it attributes too much of the observed variability in effect

sizes to sampling error. This should lead to an underestimation of  $T$  in our meta-analysis. Our sample of close replications allowed us to address this point, at least indirectly: Six meta-analyses were conducted solely on within-subjects studies. Excluding these meta-analyses did not change mean  $T$  (full set of 57 close replications:  $T = 0.09$ ; 34 close replications without within-subjects meta-analyses:  $T = 0.08$ ). This suggests that any effects of mistreating within-subjects studies as between-subjects are not large.

### 3.4.9 Conclusion and future directions

Our results on close replications show that low heterogeneity is achievable in psychology. Typically, though, it is large (see also Stanley et al., 2018; Van Erp et al., 2017), and we found little evidence that the factors that drive it are well understood. We highlighted implications of this pattern of findings for progress in psychological science and its successful application to practical problems. We hope that these examples demonstrate the usefulness of heterogeneity for thinking about research.

Such a fruitful use, of course, requires that heterogeneity is being quantified in the first place, which we found is rarely the case. We suggest  $T$  for this purpose because it has a straightforward interpretation and allows for meaningful comparisons across meta-analyses. Cohen's widely used benchmarks for effect size ( $d = 0.2/0.5/0.8$  equates to small/medium/large) map closely on the mean effect size  $\pm 1SD$  that we observed in 150 meta-analyses here ( $\delta = 0.5$ ,  $SD = 0.3$ ). Seeing  $T$  through the same lens, we might describe values of 0.2/0.35/0.5 as small/medium/large heterogeneity<sup>9</sup>. This standard might be useful to gauge heterogeneity and thus facilitate comparisons.

Power and its implications for the meaning of replication efforts (Open Science Collaboration, 2015) have played a central role in recent considerations of heterogeneity (Kenny & Judd, 2019; McShane & Böckenholt, 2014; Shrout & Rodgers, 2018; Stanley et al., 2018). Low observed heterogeneity in close replications suggests that the low replication rates in Open Science Collaboration cannot be easily dismissed as the consequence of low power, thus shifting attention back to publication bias and QRPs. Although this assessment might appear disappointing, we believe it justifies optimism because convincing counter-measures are

---

<sup>9</sup> This only applies where the effect size measure is Cohen's  $d$  (or a similar standardised mean difference). Where  $r$  is used as an effect size, halving these values appears plausible. This is because  $d \approx 2r$ , unless effect sizes are large.

available (Munafò et al., 2017). We believe that close replications will continue to play an important role in de-biasing psychological research.



## 4 Study 1b: Heterogeneity estimates and bias: A simulation study

### 4.1 Introduction

Study 1a suggests that a large amount of heterogeneity exists in the average meta-analysis. The question is, to what extent we can trust the heterogeneity estimates in Study 1a. Flexibility in the collection and analysis of data (i.e. questionable research practices, QRPs) and publication bias likely lead to upwardly-biased effect sizes in published studies (Fanelli & Ioannidis, 2013; McShane, Böckenholt, & Hansen, 2016). These in turn can give rise to upwardly biased estimations of effect size in meta-analyses (e.g. McShane et al., 2016). Heterogeneity estimates might be similarly affected. If this is the case, this will have implications for the interpretation of observed levels of heterogeneity (Stanley et al., 2018; Van Erp et al., 2017). Thus, the high levels of heterogeneity ( $T = 0.33$ ) observed in Study 1a might reflect overestimation of heterogeneity in practice, or this could be an underestimation and true heterogeneity might be even higher than what we observe.

As publication bias and QRPs cannot be studied directly, computer simulations are often used to investigate these phenomena. Previous simulations have investigated the impact of factors such as QRPs and publication bias on either effect sizes or heterogeneity estimates (for example, see Augusteijn, van Aert, & van Assen, 2019; Carter, Schönbrodt, Gervais, & Hilgard, 2019; Jackson, 2006). Carter et al. (2019) simulated an extensive range of parameters, including true effect size, heterogeneity, number of studies in a meta-analysis, publication bias and QRPs, and used a variety of meta-analytic methods to analyse these. They found that QRPs generally lead to an increase in upward bias of effect sizes using commonly utilised meta-analysis models, with QRPs exacerbating the bias caused by publication bias (Carter et al., 2019), however, they did not look at the impact of these factors on heterogeneity. Jackson (2006) investigated the impact of publication bias on heterogeneity (measured as  $\tau^2$ ). Using the DerSimonian Laird approach, Jackson found that estimates of heterogeneity can both increase and decrease under conditions of publication bias, making reliable correction for publication bias problematic (see also Ioannidis & Trikalinos, 2007). Similarly, using a Monte-Carlo simulation, Augusteijn et al. (2019) assessed the impact of publication bias on heterogeneity (measured as both  $I^2$  and using Cochrane's  $Q$ ), and found that publication bias affects heterogeneity in a non-linear way. When no non-significant studies are published, heterogeneity is overestimated, and heterogeneity becomes underestimated when publication bias reduces. In particular, when true heterogeneity was small or medium, publication bias

could create strong underestimation of heterogeneity, such that dissimilar studies can appear homogenous. Augusteijn et al. (2019) looked at publication bias only, and did not examine the impact of QRPs on heterogeneity.

The previous simulations discussed above which assessed publication bias mainly focused on one-tailed testing of hypotheses (for an exception, see Carter et al., 2019, but remember that Carter et al. did not look at the impact on heterogeneity estimates), basing this on *a priori* presumption that effects would be expected to fall in a given direction. While this might seem, upon initial consideration, to be a sensible assumption (researchers should, after all, state their hypotheses before conducting their research, and will often have assumptions underpinning these regarding the expected effect), competing theories could suggest circumstances under which a result could be expected to reverse. For example, an intervention might aim to increase positive thinking (for example, making the best of a difficult situation), and this might be expected to improve participants' mood. However, an alternative hypothesis could be that positive thinking leads to suppression of one's true emotions, so when people are struggling with low mood, suggestion to think positively could worsen their mood further, producing results opposite to the expected direction. The relatively high levels of observed heterogeneity found in study 1a suggest that it is not uncommon for the results of an individual study to run against the expected direction. In addition to this, there is reason to believe that it is not uncommon for researchers to formulate their hypotheses after examining their results (also known as HARKing: Hypothesising after the results are known, for example see Kerr, 1998), which could include hypothesising that a result counter to the usual direction was to be expected.

Meta-analysis authors can also have hypotheses or theoretical bases for judgements that impact on the selection of studies included in a meta-analysis. This can result in meta-analyses that produce biased assessments of an effect, which is perhaps reflected in the fact that meta-analyses themselves do not always resolve controversial debates, and alternative meta-analyses, with differing conclusions can be produced when researchers disagree with meta-analytic findings (Ferguson & Heene, 2012; Greenwald, 2012).

For example, the relationship between media violence (and video games in particular) and forms of aggressive behaviour has been extensively studied, and violent gaming is often implicated as a causal factor in violent crimes (Bushman & Anderson, 2015). However, meta-analyses into this effect have come to opposing conclusions (for contrasting examples, see Anderson & Bushman, 2001; Ferguson, 2015a). Some authors might argue that for some

groups of people, where violent behaviour is particularly unacceptable, gaming could provide a helpful outlet that could reduce outward expressions of aggression in other situations (for example, see Ferguson & Dyck, 2012; Ferguson, San Miguel, & Hartley, 2009). In this circumstance, we might expect effects to run counter to the usual direction. We therefore examined the effects of QRPs and publication bias in simulations based on both one-tailed and two-tailed testing.

#### 4.1.1 Aims

Here, we investigate whether observed levels of heterogeneity are trustworthy or are instead affected by publication bias and other distortions in the published literature (e.g. McShane et al., 2016; Simmons, Nelson, & Simonsohn, 2011; Sterling, 1959). Our goal is to evaluate to what extent publication bias impacts on observed levels of heterogeneity, and we extend the investigations of previous studies (for example, Augusteijn et al., 2019; Jackson, 2006) by also considering the extent to which QRPs, and one-tailed or two-tailed testing, bias observed levels of heterogeneity.

### 4.2 Method

In order to investigate the effect of publication bias and QRPs on observed heterogeneity we ran a series of computer simulations. All simulated between-subjects experiments in which the means for an experimental and a control group were contrasted. Multiple independent studies were run, published (or not), and (if published) summarised in a meta-analysis. Resulting heterogeneity estimates could then be compared against the true level of heterogeneity, which was one of the parameters varied in the simulations.

#### 4.2.1 Factors manipulated

We manipulated six factors in our simulations (for an overview, see Table 4.1): i) True heterogeneity ( $\tau$ ) had three levels: zero (absent), 0.2, and 0.4 (following Carter et al., 2019); ii) The average size of the true effect ( $\delta$ ) had four levels: zero (no effect), 0.2 (small effect), 0.5 (medium effect), and 0.8 (typically considered a large effect). Remember, where heterogeneity is present, experiments that feed into the same meta-analysis tap into different effect sizes. In this case,  $\delta$  describes the mean of this effect size distribution (e.g. in Figure 3.2,  $\delta = 0.45$ ). ; iii) Number of studies per meta-analysis ( $k$ ) had four levels: 10, 30, 60 and 100; iv) We conducted the simulations under conditions of either one-tailed testing or two-tailed testing for the

simulated studies; v) We manipulated the strength of publication bias, with three levels: zero (no publication bias), 40% publication bias, and 80% (following Carter et al., preprint). Statistically significant findings were always published (and therefore always entered the meta-analysis). If an experiment returned a non-significant result ( $p > .05$ ), the probability that it would be purged ranged from zero (no publication bias) to 80%. Under publication bias, the simulation kept running until the target number of  $k$  published studies was reached for a given meta-analysis; vi) QRP environment, which is described in more detail below.

Our final manipulated factor requires some additional explanation as this concerns the type and prevalence of QRPs applied to the experiments. Following Carter, Schönbrodt, Gervais, and Hilgard (2019), we considered four types of QRPs. i) Optional dependent variables: Researchers in the simulated experiment use two dependent variables. At the population level, these variables always correlated  $r = .8$ . Researchers only publish the finding that results in the smaller  $p$  value. ii) Optional stopping: Researchers regularly peek at their results. If the results is non-significant, they add participants (at 10% of the starting sample size), and continue this process until they either reach a statistically significant finding, or hit the maximum number of participants (set as either five times the starting sample size, or at an absolute maximum of 200 per condition). iii) Optional moderator: The sex of all participants in the experiment was decided at random ( $p = .5$ ). Researchers analyse results for females only, males only, and the whole sample. They only publish the finding that results in the smallest  $p$  value. iv) Optional outlier removal: Researchers run analyses on all data, and on data with outliers (unsigned  $z \geq 2$ ) removed. They only publish the finding that results in the smallest  $p$  value.

A survey of researchers in psychology (John, Loewenstein, & Prelec, 2012) suggested that these four QRPs are all prevalent and often regarded as unproblematic. We then used these four QRP types to simulate three research environments, following Carter, Schönbrodt, Gervais, and Hilgard (2019): no QRPs, medium QRPs, and high QRPs. In the no QRP environment, no QRPs were used. In the medium QRP environment, 30% of researchers did not engage in QRPs, 50% of researchers used both optional dependent variables and optional stopping, and 20% of researchers used all forms of QRPs. For the high QRPs environment, these percentages were 10%, 40%, and 50%, respectively.

An overview of all manipulated factors and their levels is provided in Table 4.1. This resulted in 864 unique combinations of six fully crossed factors. For each, 2,000 meta-analyses were simulated. All meta-analyses were conducted with three meta-analysis models in Metafor in R: DerSimonian Laird, Hunter-Schmidt, and Paule-Mandel (see section 2.1.5 for details). This

allowed us to compare how well these models estimate heterogeneity under various conditions.

Table 4.1 Simulation Parameters for Study 1b.

Experimental Factors	Levels
One- or two-tailed testing	One, two
Size of true population effect ( $\delta$ )	0, 0.2, 0.5, 0.8
True heterogeneity ( $\tau$ )	0, 0.2, 0.4
Number of studies per meta-analysis ( $k$ )	10, 30, 60, 100
% of studies affected by publication bias	0%, 40%, 80%
QRP environment	None, medium, high

#### 4.2.2 Technical details

An annotated version of the R script used to run the simulations is available in Appendix B. Here, we describe some of the key details.

The simulation started by drawing random samples from the control and experimental populations. Population variances were always one, and the difference between their population means ( $\mu$ ) reflected two factors in our simulation, the specified level of effect size ( $\delta$ ) and heterogeneity ( $\tau$ ). When  $\tau = 0$ , the difference in population means always equalled the effect size; for example,  $\mu_{\text{control}} = 0$ ,  $\mu_{\text{experimental}} = 0.5$  for  $\delta = 0.5$ . When  $\tau > 0$ , the population effect size for a given experiment was drawn at random from a distribution with a mean of  $\delta$ , and a standard deviation of  $\tau$ . For example, imagine that  $\delta = 0.5$  and  $\tau = 0.2$ . The majority (95%) of observed effect sizes ( $d$ ) would be drawn at random from a range of effect sizes from 0.11 to 0.89.

We aimed for a realistic distribution of sample sizes in the simulations. We therefore collated all individual study sample sizes of four (two per group) or more from the 150 meta-analyses examined in study 1a. As these were total sample sizes, each of these totals was divided by two to provide an approximation of each sample group size. This information was imported into R, which then selected a sample size from this list (totalling 7228 sample sizes) at random. The minimum sample size was set at 2 per group. Where optional stopping was implemented, the starting sample size was the sample size at which the first peek at the data was taken.

For each experiment, an unbiased estimate for Cohen's  $d$  (Lakens, 2013) and sample size were fed into its respective meta-analysis. In very rare cases, simulations with QRPs generated unrealistically large effect sizes. To maintain the realism in our simulations, we decided to discard extreme results; we arbitrarily set the threshold at  $|d| = 5$ . Meta-analyses were conducted in Metafor, and relied on the same models as in study 1a to compute  $d$  and  $T$  as estimates for  $\delta$  and  $\tau$ . As with study 1a, we focus mainly on the DL estimator, with results from the HS and PM estimators in Appendix B.

For the 2,000 simulations in each unique factor combination, we summarised results using  $T_{bias}$  to capture systematic bias in the estimation of  $\tau$ .  $T_{bias}$  was calculated as estimated  $T$  minus true heterogeneity ( $\tau$ ) using the following formula:

$$T_{bias} = T - \tau$$

Equation 4.1

### 4.3 Results

#### 4.3.1 Choice of heterogeneity estimator

We first consider our three estimators of heterogeneity, namely DerSimonian Laird (DL), Hunter Schmidt (HS), and Paule-Mandel (PM). We considered both  $T_{bias}$  and unsigned error in  $T$  in order to evaluate which of the meta-analysis estimators produced the least amount of bias and error across all levels of the included factors. We considered it important to assess both unsigned error in  $T$  in addition to  $T_{bias}$ , as low  $T_{bias}$  across conditions can mask that overestimation and underestimation can cancel each other out. For example, consider the middle, left hand panel in both Figure 4.1 and Figure 4.2, which show the effect of Effect size on  $T_{bias}$  and unsigned error in  $T$ , respectively. For  $T_{bias}$ , it appears that Effect size does not affect  $T$  to any great extent, but the unsigned error in  $T$  shows a different pattern of results, namely that low effect sizes seem to create greater error in  $T$  than larger effect sizes.

Figure 4.1 and Figure 4.2 show the main effect of all simulated factors on  $T_{bias}$  and unsigned error in  $T$  respectively. Across all factor levels, DL shows the least amount of  $T_{bias}$  for 12 out of 19 of the factor combinations (and never shows the highest amount of  $T_{bias}$  for any of the factor combinations). That is, the estimates of heterogeneity obtained from simulation conditions using this model seem to be closer to the true level of heterogeneity ( $T_{bias} = 0$ ) for the DL estimator than for either of the other meta-analysis estimators. Similarly, for unsigned

error in  $T$ , DL showed the least amount of error across 17 of the 19 factor combinations.

Overall DL seems to produce the least amount of bias or error in  $T$ , and all subsequent analyses therefore focus on results from the DL estimator only.

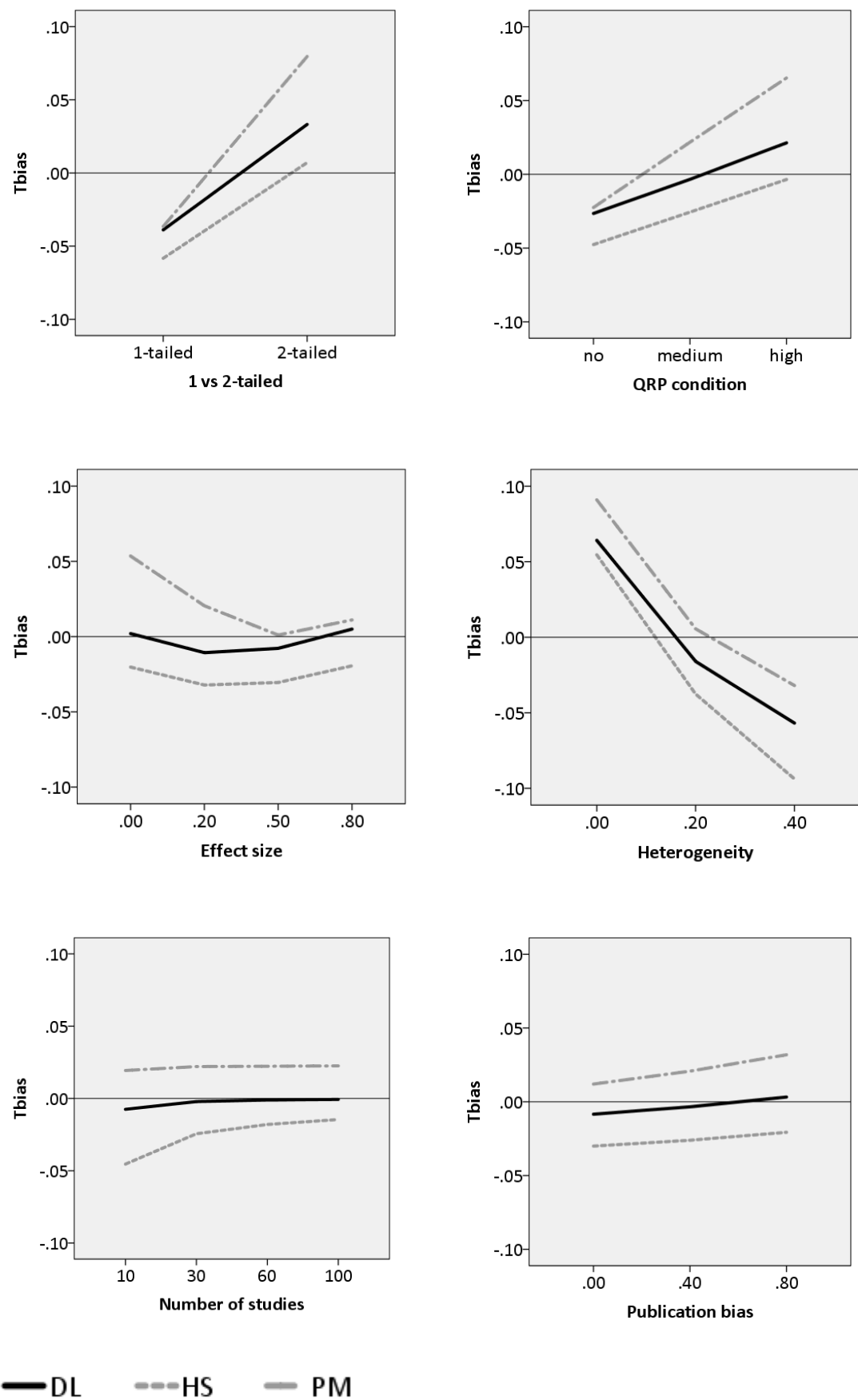


Figure 4.1  $T_{bias}$  across all factor levels.



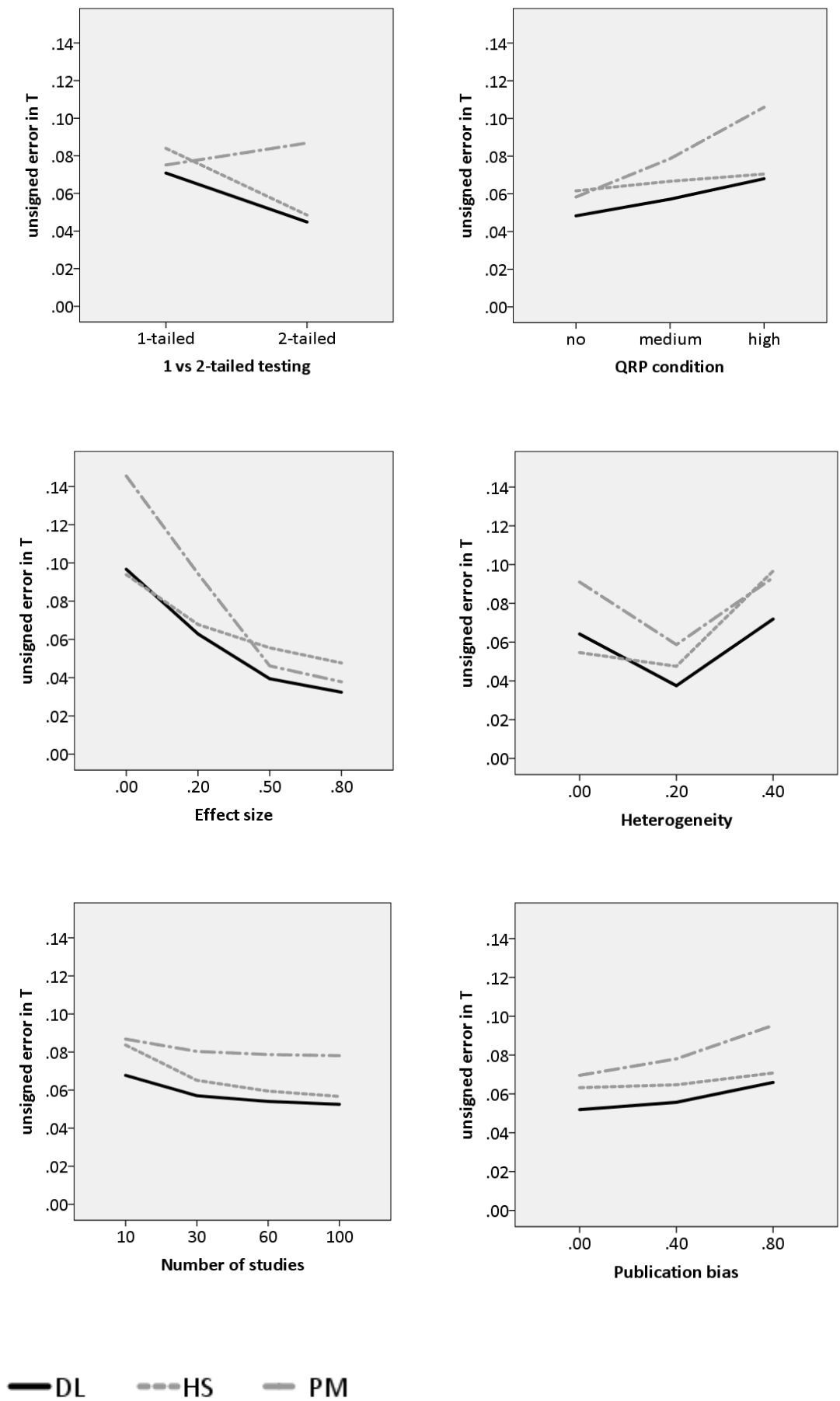


Figure 4.2 unsigned error in  $T$  across all factor levels.

#### 4.3.2 Bias in heterogeneity estimates

Overall  $T_{\text{bias}}$  across all factor conditions was -0.003. However, as Figure 4.3 shows, although there was no bias overall, it is clear that substantial over- and under-estimation of  $T$  is occurring within the different simulation conditions. We therefore explore the different factor combinations to determine under which conditions this bias occurs.

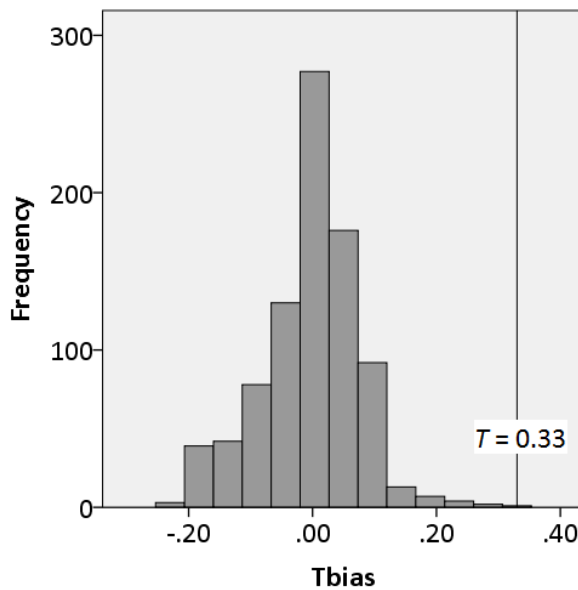


Figure 4.3 Distribution of  $T_{\text{bias}}$  across all simulation conditions

To cut through the complexity of results, we first ran a 6-factorial ANOVA on the data. This used the six manipulated factors (see Table 4.1). The ANOVA was run only in order to compute  $\eta^2$  for the manipulated factors and their interactions –  $\eta^2$  has no intrinsic meaning, but allows us to get an overview of the relative importance of each of the factors and their interactions. The results presented are therefore descriptive rather than inferential. Results on the six main effects and on all two-way interaction effects can be found in Table 4.2. (All higher order interactions proved trivial and are therefore omitted.) From Table 4.2 it is apparent that three main effects were particularly strong, namely the effects of one-tailed versus two-tailed testing, QRPs, and true heterogeneity. Note that two of these effects are due to simulation parameters that we consider in relation to heterogeneity for the first time in the current study (one- versus two- tailed testing, and QRPs).

For one-tailed versus two-tailed testing, one-tailed testing led to a slight underestimation of heterogeneity ( $M = -0.04$ ), whereas two-tailed testing led to a slight overestimation of heterogeneity ( $M = 0.03$ ). As one- versus two-tailed testing affects the predominant direction of  $T_{\text{bias}}$ , we report the remaining results separately for one-tailed and two-tailed testing.

In terms of QRP condition, for one-tailed testing, heterogeneity was, on average, slightly underestimated (for no QRPs,  $M = -0.06$ ; for medium QRPs,  $M = -0.04$ ; for high QRPs,  $M = -0.02$ ). For two-tailed testing, heterogeneity was slightly overestimated (for No QRPs,  $M = 0.01$ ; for medium QRPs,  $M = 0.03$ ; for high QRPs,  $M = 0.06$ ).

In terms of the impact of heterogeneity, for one-tailed testing,  $\tau = 0$  led to slight overestimation of heterogeneity ( $M = 0.05$ ), whereas on average, higher levels of heterogeneity led to underestimation of heterogeneity, and this was greater for  $\tau = 0.4$  ( $M = -0.12$ ) than for  $\tau = 0.2$  ( $M = -0.05$ ). For two-tailed testing, heterogeneity was, on average, slightly overestimated (for  $\tau = 0$ ,  $M = 0.08$ ; for  $\tau = 0.2$ ,  $M = 0.02$ ; for  $\tau = 0.4$ ,  $M = 0.003$ ).

Only one two-way interaction proved of magnitude to be of particular interest, that is 1 vs 2 tailed  $\times \delta$  (see Figure 4.4). This shows that heterogeneity is often underestimated for one-tailed testing ( $M = -0.07$ ), whereas for two-tailed testing heterogeneity was overestimated ( $M = 0.04$ ).

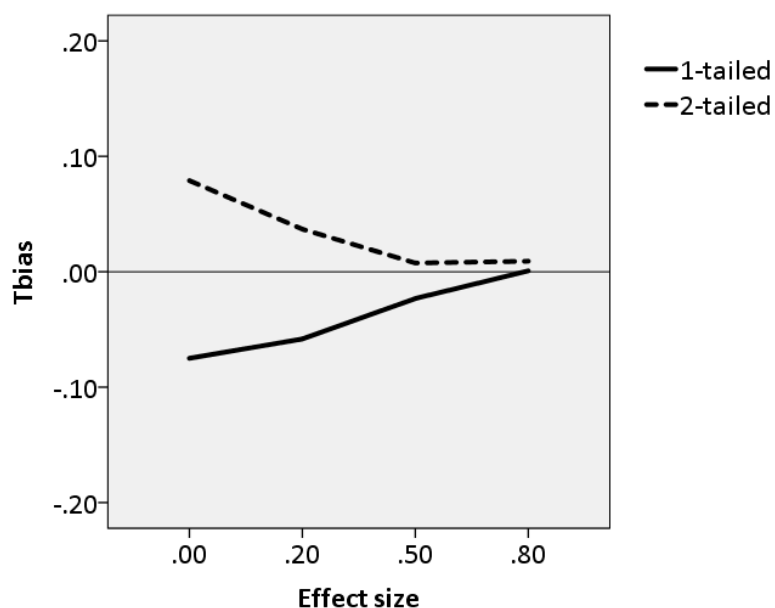


Figure 4.4  $T_{\text{bias}}$  for 1-tailed and 2-tailed testing, across different levels of effect size

Table 4.2 Initial effects of all simulation parameters on  $T_{\text{bias}}$ .

	Eta squared ( $\eta^2$ )
True heterogeneity ( $\tau$ )	.415
One versus two tailed	.213
QRP condition (No, medium, high)	.063
True effect size ( $\delta$ )	.007
Publication bias (PB)	.004
Number of studies ( $k$ )	.001
1 vs 2-tailed x $\delta$	.133
1 vs 2-tailed x $\tau$	.052
$\delta$ x $\tau$	.034
$\delta$ x PB	.017
QRP x $\tau$	.011
QRP x $\delta$	.006
T x $k$	.004
1 vs 2-tailed x QRP	.002
1 vs 2-tailed x PB	.002
1 vs 2-tailed x $k$	.001
$\delta$ x $k$	.001
T x PB	.001
QRP x $k$	.000
QRP x PB	.000
$k$ x PB	.000

*Note:* Effect size calculations were based on  $N = 2$  per cell, where each  $N$  represents the average of 1,000 simulations. Effects are presented in order of magnitude (within main effects and interaction effects).

#### 4.4 Discussion

Publication bias and QRPs are likely to bias the published record of psychology (Ferguson & Brannick, 2012; John et al., 2012; Kühberger et al., 2014; LeBel & Peters, 2011; Levine et al., 2009; Sterling, 1959). Our computer simulations suggest that one-tailed versus two-tailed testing has a strong impact on bias in heterogeneity estimates, which highlights the value of including both types of testing in our simulations. In a one-tailed testing environment, these biases appear to lead to the underestimation of true heterogeneity, in some cases quite

significantly (mean underestimation  $T_{\text{bias}} = -0.12$ ). However, under conditions of two-tailed testing these biases inflate true heterogeneity (especially where true effect size and true heterogeneity are small), but to an extent that is small in relation to actual observed heterogeneity.

The main goal of our simulation was to provide perspective on observed levels of heterogeneity observed in Study 1a, to determine whether these might be providing a biased estimate of true heterogeneity in meta-analyses. Our average observed  $T$  in Study 1a (0.33) never occurred in the simulation study (the maximum level of overestimation of  $T$  was 0.31, and the maximum underestimation was -0.23), meaning that the heterogeneity we observed in Study 1a cannot be wholly due to bias. In fact, considering  $T$  of half of the magnitude observed in Study 1a, even this is rare in our simulations (see Figure 4.3; this magnitude of bias occurred in only 1.7% of our 864 simulation conditions). In sum, there is no reason to believe that the conclusion from Study 1a that heterogeneity tends to be very large is mistaken.

In addition to this, in Study 1a, one of the key findings was a strong positive correlation between meta-analyses' effect size estimates and  $T$  ( $r_{\text{win}} = .49$ ). Our simulations show that this is not an artefact caused by a positive relationship between  $\delta$  and  $T_{\text{bias}}$ , because  $\delta$  has no strong effect on  $T_{\text{bias}}$  (see Table 4.2). Similarly, in study 1a we observed a positive relationship between  $k$  and  $T$  ( $r_{\text{win}} = .23$ ). Potentially, this could be an artefact in that this could be due to bias in  $T$ . However, the relationship between  $k$  and  $T$  observed in Study 1a seems unlikely to be an artefact, because  $k$  had no strong effect on  $T_{\text{bias}}$ .

## 5 Study 2: Peaks of psychology: The very largest effects

### 5.1 Introduction

As the science of behavioural phenomena, psychology seeks to understand human behaviour and mental processes, and seeks to understand and solve problems by developing valuable and suitable practical applications. Undoubtedly, both of these aims are advanced when pertinent effects are large in magnitude. Study 2 seeks to identify the strongest effects in psychology.

Where a hypothesis seeks to explain or predict a given phenomenon, a large effect size will be indicative of a particularly powerful explanation for this effect. Take the wellbeing of adolescents as an example. An intervention that increased the number of hours of sleep obtained per night could lead to greater wellbeing in English adolescents, compared to those who did not receive the intervention (cf. Gireesh, Das, & Viner, 2018). This finding would have different meaning depending on the magnitude of the effect. For example, imagine that Paul (15 years old) has low scores on a measure of wellbeing, and sleeps for an average of only 5 hours per night. What help can we suggest to improve Paul's wellbeing? If the effect of the sleep intervention was not particularly strong (e.g.  $d = 0.20$ ), this might suggest that the sleep intervention could have a small effect on Paul's wellbeing, whereas if the effect were very strong (e.g.  $d = 1.50$ ) this might suggest that the sleep intervention could lead to a substantial improvement in his wellbeing. Addressing sleep difficulties could therefore have a different impact on adolescent wellbeing, from proving somewhat helpful, to alleviating the problem. If the effect is strong, we have a clear idea of how to help Paul, which is not the case if the effect is weak.

Furthermore, when an effect shows a low proportion of effect sizes that go against the expected direction (*Flop Index (FI)*), this indicates reliability of this explanation. In our example, this would indicate that the effect of improved sleep on wellbeing in adolescents is particularly consistent. Consistency across studies examining the same effect would suggest that the independent variable has the same effect across different samples, methods, and circumstances. Continuing our example above, in this case consistency would suggest that under low heterogeneity and low *FI*, this study would be relevant to Paul if he were to live in Turkey rather than in England. If a hypothesis relates to the nature of an observed effect, a large effect and low *FI* is indicative of this effect being reliably observed or evoked, and this should, in turn, assist efforts to explain it.

In study 1a, we found that the average effect size in psychology is medium ( $d = 0.47$ ), similar to that found by several large-scale replications and other surveys of meta-analyses (for example, see Bosco, Aguinis, Singh, Field, & Pierce, 2015; Brysbaert, 2019; Gignac & Szodorai, 2016; Stanley et al., 2018). Under average conditions of heterogeneity, we found in meta-analyses in study 1a (average  $T = 0.33$ ), this effect size would result in a percentage of effects ( $FI = 9\%$ ) running against the expected direction. When heterogeneity is high, it therefore becomes difficult to predict the results of the next study with any degree of certainty. This begs the question – under high heterogeneity, which effects might be more consistent, at least in terms of the direction of the effect? Although the result of a practical application (which can also be thought of as the ‘next study’) will always have a degree of uncertainty, all else being equal, larger mean effect size relating to a given phenomenon should mean that the next study is more likely to have a large effect size. In addition to this, with average levels of heterogeneity, larger effect sizes will minimise the chance of the effect of the next study running against the expected direction, and therefore maximise the chances of consistency in the direction of the effect.

Examination of very-large-effects is of particular interest as this can suggest that a result is likely to have particular promise for explanation of effects and for the implementation of practical interventions, and is more likely to be consistent in direction. However, different sub-disciplines in psychology could show differences in the likelihood of meta-analyses being included in our very-large-effects sample. This could be informative with regards to explanation of effects (for example whether a small number of powerful causes exist to explain phenomena, versus multiple, smaller causes). It could also be informative with regards to strengths in research methods. If very-large-effects in certain disciplines appear frequent due to particularly successful research methods, other disciplines could learn from that. We therefore thought to consider whether different sub-disciplines in psychology had equal representation in the very-large-effects sample within this time period. The absence of large effect sizes from a specific subfield could suggest that further investigation might elucidate the reasons for this.

The key aims of Study 2 are: i) to establish the strongest effects that have already been determined in psychology and; ii) which areas of psychology these refer to. This study therefore aimed to examine the largest effect sizes in psychology, based on recently published meta-analyses, and to examine their associated levels of heterogeneity. This will indicate

which effects might be most likely to provide strong explanations for phenomena, and which effects might be most likely to result in successful practical applications.

## **5.2 Method**

### **5.2.1 Study search and selection strategy**

We intended to assess the strongest effects in psychology, in order to determine in which areas and under which conditions these are observed. We therefore aimed to obtain approximately the 100 largest reported meta-analysis average effect sizes (henceforth, ‘very-large-effects’).

We searched PsycINFO, journals only, for “meta-analy\*” in the abstract field, in December 2017. For feasibility reasons, we restricted the search to the last 5 years. This resulted in a total of 10,871 meta-analyses (see Figure 5.1). We then used the mean and distribution of effect sizes from study 1a to determine a minimum effect size threshold for inclusion, as 100 meta-analyses equates to approximately the largest 1% of effect sizes. This resulted in a minimum Cohen’s  $d$  of 1.65, which was also converted to Pearson’s  $r$  of 0.64, and Log Odds Ratio (Log OR) of 3.00, which equates to an Odds Ratio (OR) of 20.09, following the calculations outlined in section 2.1.1.

#### **5.2.1.1 Inclusion criteria**

Meta-analyses were screened (full text) and included if they met all of the following criteria: i) The analysed effects were described as standardised mean differences (Cohen’s  $d$ , Hedges  $g$ ), correlations (Pearson’s  $r$  or Fisher’s  $Z$ ), or Log Odds Ratios or Odds Ratios. ii) The mean effect met the minimum threshold for effect size. iii) For practical reasons, the full article had to be available in English.

In this way, we identified 43 meta-analyses with very-large-effects (thus falling short of the intended target of 100). These were categorised using the sub-disciplines utilised by PsycINFO (see Appendix C for a list of these categories). A further 100 meta-analyses were sampled at random from the original PsycINFO search results. This provided a comparison (control) sample for the body of meta-analyses with very-large-effects.



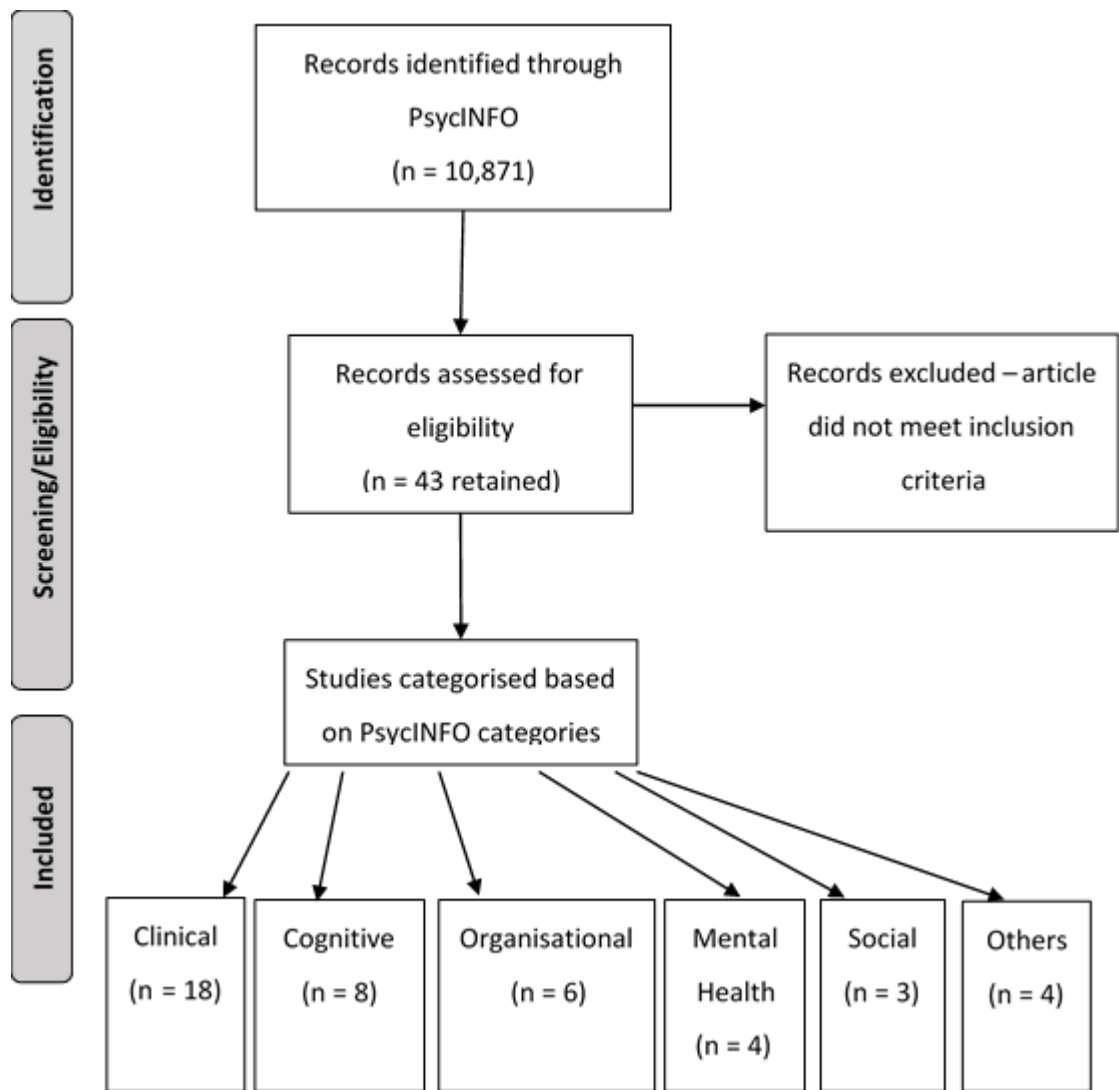


Figure 5.1 Sampling of meta-analyses

### 5.2.2 Data extraction and analysis

Following a similar protocol to study 1a, where the results of more than one meta-analysis were reported, the one including the largest number of studies was extracted. If multiple meta-analyses included the same number of studies, the first was used. A measure of heterogeneity was also recorded where this was reported.

We also calculated the mean effect size and heterogeneity ( $T$ ), using Metafor (see section 2.1.5 for more information), for all meta-analyses which reported sufficient information ( $k = 23$  for very-large-effects sample;  $k = 43$  for control sample). Following study 1a, results are reported from the DL estimator, with results of the other estimators of heterogeneity (HS and PM) reported in Appendix C.

### 5.2.3 Error and the meaning of very-large-effects

As our very-large-effect sample is selected on the basis of effect size, it is important to consider some of the potential sources of error that could inflate effect size. Consequently, these errors might be overrepresented among the very-large-effect meta-analyses, which would obviously be undesirable.

The effect size obtained from a meta-analysis will reflect both the true effect size and an unknown amount of random sampling error. This random error impacts on our sample of very-large-effects, as this is an example of extreme data. As the relationship between the estimated effect size (that which we observe in our very-large-effects sample) and the true effect is not a perfect correlation, the true effect size could often be less strong than our estimate due to regression to the mean. Where multiple meta-analyses have been conducted on the same topic, including different primary studies, that all show a large summary effect, this would alleviate these concerns.

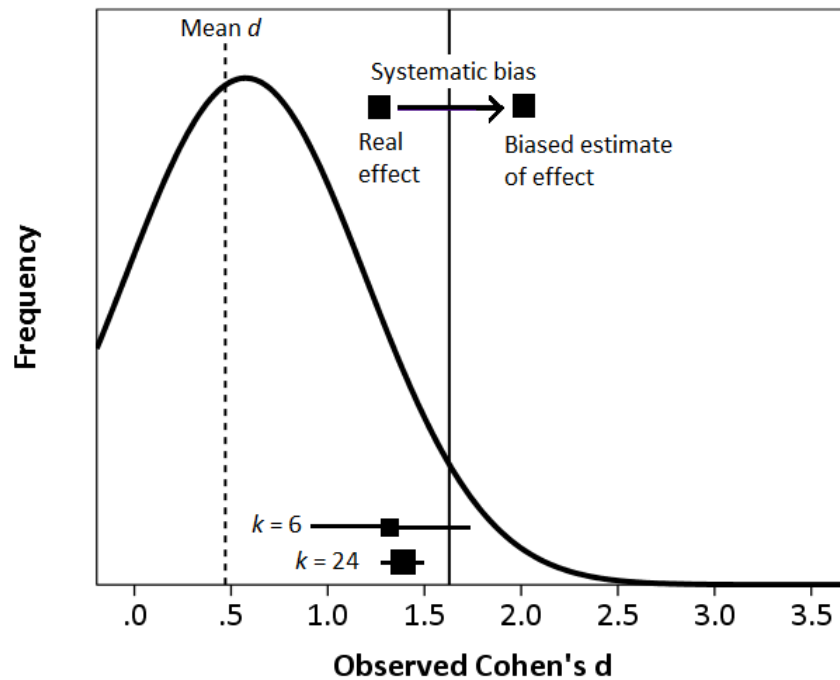


Figure 5.2 Effects of error and bias on inclusion of meta-analyses in our sample. Vertical line at  $d = 1.65$  is our effect size cut off. Small  $k$  meta-analyses ( $k = 6$ ) can end up in our very-large-effects sample through error, whereas this is less likely for larger  $k$  meta-analyses ( $k = 24$ ). Systematic bias (including QRPs and publication bias) can inflate true effect sizes so that they can end up in our very-large-effect sample.

Sample size (either  $N$  or  $k$ ) can also be a source of random error, and here we consider  $k$  (number of studies in a meta-analysis). *Ceteris paribus*, where a meta-analysis contains only a small number of studies, the precision of the effect size estimate will be lower than if a meta-analysis contains a large number of studies. Where a meta-analysis contains only a small number of studies, it is therefore more likely that the random error pushes its effect size over the selection threshold for very-large-effects (see Figure 5.2).

Meta-analyses also contain systematic forms of error, such as questionable research practices (QRPs) and publication bias (see Study 1a/b). These sources of error are present in the literature, rather than being caused by the meta-analyst. In addition to this, systematic bias can occur due to choice of studies to include, or choice of outcome measure considered by meta-analytic authors, which can also affect whether a meta-analysis could be included in our sample (see Figure 5.2). It is unfortunately rare for meta-analyses to include their own assessment of bias, and this can be difficult for readers to objectively assess, particularly as one cannot be sufficiently informed about all phenomena being subject to meta-analysis in order to assess this for a wide range of topics. However, arguably, the concepts of bias and

meta-analytic quality could be linked, and we therefore assess meta-analytic quality as described below.

#### 5.2.3.1 *Assessment of the quality of meta-analyses*

Previous studies have shown that the reporting standards and methodological rigour in meta-analyses are variable (Aytug, Rothstein, Zhou, & Kern, 2012; Dieckmann et al., 2009), and that quality of reporting can be related to the prevalence and magnitude of errors in the reporting of statistical results (Bakker & Wicherts, 2011). A reliable and valid measure of the methodological quality of systematic reviews already exists to assess this, in the form of the ‘assessment of multiple systematic reviews’ (AMSTAR, Kang et al., 2012; Sharif, Janjua-Sharif, Ali, & Ahmed, 2013; Shea et al., 2007; Shea et al., 2009). AMSTAR is composed of an 11-component tool that assesses different components of systematic reviews and meta-analyses for methodological quality and bias. For example, whether the research design was specified *a priori*, the comprehensiveness of the literature search conducted, inclusion of the characteristics of primary studies, and assessment of quality of primary studies included.

Conducting a full AMSTAR rating on each of our included meta-analyses was not feasible due to time constraints, and we therefore selected four components from the AMSTAR for which data were particularly easy to extract and for which determination of meeting the criteria for these items was objective: inclusion of an assessment of heterogeneity; inclusion of an assessment of publication bias; inclusion of grey literature. To capture the fourth component, we also created a score for reporting of the primary studies included in the meta-analysis, awarding 0.34 score for each of the following: inclusion of a list of primary studies; inclusion of effect size; inclusion of sample size (these scores were calculated in this way as reporting of primary studies in this way forms a single question on the AMSTAR). Combining these scores generated a quality score for each meta-analysis ranging from 0 (low quality) to 4 (high quality).

As discussed previously (see study 1a), errors in reporting could add error variance to meta-analyses, and therefore inflate observed levels of heterogeneity. In addition to this, as indicated above, error could lead to overrepresentation of poor quality meta-analyses in our very-large-effects sample (see Figure 5.2). Our assessment of meta-analytic quality attempts to capture this here.

### 5.2.3.2 *Assessment of triviality and informative nature of effects*

Our sample could include meta-analyses that report effects that might appear to be of little interest, either in terms of what is added to conceptual knowledge and understanding, or the impact of an intervention or practical implication. For example, a strong relationship between two measures of the same concept (such as intelligence) would seem hardly informative. While obtaining a strong relationship might be reassuring for studying the effect in question, it does not add to new conceptual knowledge.

In addition to this, a large effect size alone might not be particularly informative if a meta-analysis also has particularly high heterogeneity. For example, communication therapy aimed at helping autistic people to communicate might show a very large effect size, but if this is accompanied by large heterogeneity, it would be important to understand the sources of this heterogeneity, and whether the therapy could in fact be detrimental for some people. In study 1a, we were able to predict levels of heterogeneity based on the effect size from any given meta-analysis ( $T = 0.18 + 0.30d$ ). We use this regression equation to predict expected levels of heterogeneity for each meta-analysis in the current study, then calculate residuals (actual minus predicted) to give an estimate of how much higher (or lower) heterogeneity is compared to this expected amount. Following study 1a, we also calculated the *Flop Index* (*FI*) for each of the very-large-effects. *FI* tells us the proportion of the results from subsequent studies into the effect that we would expect to fall in the opposite direction to the effect found by the relevant meta-analysis. As we are looking only at very-large-effects, *FI* will be driven primarily by the level of heterogeneity in each meta-analysis, and so presenting this measure in addition to heterogeneity might appear unnecessary here. However, we have included *FI* as this provides an at-a-glance assessment of consistency in direction of the effect.

We therefore use our own, informal judgment to label each meta-analysis as trivial or informative based on the nature of the effect studied, the amount of heterogeneity relative to that predicted, and the *Flop Index* for the effect.

## 5.3 Results

### 5.3.1 Our samples

We intended to sample the top 1% of effect sizes from meta-analyses in psychology. We determined the cut off for the lowest effect sizes to be included on a calculation based on the mean and distribution of effect sizes from study 1a. Study 1a included only meta-analyses from

cognitive, organisational, and social psychology which described effects as standardised mean differences (Cohen's  $d$ , Hedge's  $g$ ) or correlations (Pearson's  $r$ , or Fisher's  $Z$ ). For the current study, we wanted to include the full range of sub-disciplines in psychology, and recognised that some of these sub-disciplines used alternative effect sizes, in particular, Log Odds Ratios (Log OR) or Odds Ratios (OR).

We obtained only 43 meta-analyses for our sample using this method, a smaller sample than intended (top 0.4%, rather than the top 1%, see Figure 5.1). None of the included meta-analyses utilised OR or Log OR as the reported effect size. This avoids any difficulties regarding comparability of  $T$  with other effect size measures.

Following study 1a, we recalculated heterogeneity using Metafor for all studies which reported sufficient information for this analysis (effect sizes and sample sizes for each of the primary studies included in the meta-analysis). Heterogeneity could be calculated in this way for 23 meta-analyses in the very-large-effect size sample, and 43 meta-analyses in the control sample. One outlier (Cohen's  $d > 8$ ) was removed from the sample with very-large-effect sizes before conducting heterogeneity analyses.

Table 5.1 shows descriptive statistics for the meta-analyses in each sub-discipline, for both the sample of meta-analyses with large effects and the control sample. As the expected counts for many cells were under 5, thus violating assumptions of  $\chi^2$ , a permutation-based  $\chi^2$  test was conducted. A constrained permutation test was conducted using R package *vegan* (Oksanen et al., 2007), which placed constraints on the columns in the  $\chi^2$  table. This means that the sub-discipline totals (count for very-large-effects plus count for control for each column in Table 5.1) were fixed. This is an appropriate choice since it is more likely that any meta-analysis drawn from the population (10,871 meta-analyses conducted in the last 5 years and referenced by psycINFO) would be from some disciplines (e.g. clinical psychology) than others, and this method adjusts for this. 10,000 permutations were run, allowing shuffling of both rows and columns in the  $\chi^2$  table. The mean  $\chi^2$  value for these permutations was 41.45. Percentiles were used to determine the probability ( $p$ -value), assuming a 1-tailed distribution ( $\chi^2$  is always a 1-tailed distribution as it is asking whether the value found is different to that under chance.) This resulted in a significant result ( $p < .05$ ). This indicates that there is a statistically significant difference between the relative frequencies of meta-analyses in the very-large-effects sample and the control sample. Consequently, there is evidence to suggest that meta-analyses from some sub-disciplines are more likely to find very-large-effects than those from other sub-disciplines. Meta-analyses in clinical and social psychology were

represented less frequently in the very-large-effect meta-analyses than in the general sample. Those in cognitive or organisational psychology were represented more frequently.

The topics resulting in the 43 very-large-effect sizes can be found in Table 5.4. In brief, several of the meta-analyses with the largest effects in clinical psychology considered different treatments for mental health conditions, including depression, anxiety, obsessive-compulsive disorder (OCD), post-traumatic stress disorder (PTSD), and psychosis. In cognitive psychology, several meta-analyses considered the improvement of IQ scores in children due to adoption or due to training, as well as the relationship between children's word reading and reading comprehension, the association between reading comprehension and vocabulary knowledge in children learning a second language, interventions to improve grammar acquisition in children learning a second language, and differences in communication attitudes in children who stutter and those who do not. In organisational psychology, work relationships, work behaviour and leadership were all among the largest effects. See Table 5.4 for further details of other sub-disciplines and topics.

We considered the possibility that particularly large effects could be found in correlational studies that were exploring constructs that were deemed to be closely related, and that these effects might be less informative than large effects from experimental studies. We therefore sought to code meta-analyses as based on either experimental or quasi-experimental in nature, by examining the study designs discussed in the meta-analyses. Where meta-analyses included primary studies based on both types of design, the most commonly included design was assigned to the meta-analysis. In an exploratory analysis, we tested whether there was any relationship between the likelihood of a meta-analysis having a large effect and whether the meta-analysis was based on experimental or quasi-experimental studies. We found no significant association ( $\chi^2 (1) = 3.15, p = .076$ ). We also thought to test whether there was any difference in effect size between experimental (Mean  $d = 2.19$ ) and quasi-experimental (Mean  $d = 2.98$ ) studies in our very-large-effects sample. We found no significant difference ( $t(40) = 1.60, p = .117$ ). Thus, there is no strong evidence to suggest that quasi-experimental studies are over-represented in our very-large-effects sample, and that the very-large effects sample might be characterised as encompassing less informative effects than the control sample.

Table 5.1 Number of meta-analyses for each sub-discipline (with percentage of sample)

Discipline / Sample	Cognitive	Psychometrics	Experimental	Developmental	Social	Personality	Clinical	Mental health	Educational	Organisational
Very-large-effects	8 (19%)	2 (5%)	1 (2%)	0	3 (7%)	0	18 (42%)	4 (9%)	1 (2%)	6 (14%)
Control	2 (2%)	5 (5%)	1 (1%)	2 (2%)	13 (13%)	4 (4%)	59 (59%)	7 (7%)	0	7 (7%)

Table 5.2 Means (and standard deviations) for meta-analysis quality and heterogeneity ( $T$ ). For  $d$  and  $T$ ,  $k = 23$  for the large effects sample, and  $k = 43$  for the control sample.

Sample	Quality score	$d$	$T$
Very-large-effects	2.43 (0.98)	2.28 (1.55)	0.79 (0.43)
Control sample	2.22 (1.01)	0.43 (0.42)	0.33 (0.31)

Table 5.3 Correlations for exploratory analyses on what drives heterogeneity.  $k = 22$  for large sample (one outlier removed),  $k = 43$  for control sample.

Sample	$T$ and $d$	$T$ and $k$	$T$ and quality	$d$ and $k$
Very-large-effects	$r = .483, p = .023$	$r = -.088, p = .698$	$r = -.364, p = .095$	$r = -.134, p = .553$
Control sample	$r = .610, p < .001$	$r = -.116, p = .461$	$r = -.252, p = .103$	$r = -.023, p = .882$
Total sample	$r = .697, p < .001$	$r = -.087, p = .492$	$r = -.264, p = .033$	$r = -.040, p = .753$



Table 5.4 Topics with the largest effect sizes in different sub-disciplines. Note that topics were grouped based on examination of similar effects.

Topic	Effect size ( <i>d</i> )	Heterogeneity ( <i>I</i> )	Number of studies ( <i>k</i> )	Experimental or quasi-experimental	Quality (0-4)	Residual <i>T</i> (actual minus predicted)	Flop Index ( <i>FI</i> ) %	Trivial/important/questionable effect	References
Clinical									
<b>TMS (Transcranial Magnetic Stimulation) for PTSD and depression</b>									
rTMS reduces PTSD symptoms	1.65	0.00	5	Exp	3	-0.68	0.0	I	(Berlim & Van den Eynde, 2014)
rTMS reduces PTSD symptoms	2.70	n/a	13	Exp	2.67	-	-	I	(Yan, Xie, Zheng, Zou, & Wang, 2017)
DTMS improves depression scores	2.04	n/a	9	Exp	2.33	-	-	I	(Kedzior, Gellersen, Brachetti, & Berlim, 2015)
<b>CBT for OCD</b>									
Intensive CBT improves OCD symptoms in adults	2.43	0.17	17	Exp	3	-0.74	0.0	I	(Jonsson, Kristensen, & Arendt, 2015)

Intensive CBT improves OCD in adolescents	1.86	0.69	55	Exp	3.67	-0.05	0.0	I	(Rosa-Alcázar et al., 2015)
Intensive CBT improves OCD in children	1.74	0.90	11	Exp	4	0.20	2.7	I	(Sánchez-Meca, Rosa- Alcázar, Iniesta- Sepúlveda, & Rosa- Alcázar, 2014)
<b>Treatment of anxiety and depression</b>									
Medications improve symptoms in anxiety disorders	2.02	n/a	206	Quasi	2.33	-	-	I	(Bandelow et al., 2015)
Meta-cognitive therapy improves symptoms of anxiety and depression	2.00	0.49	16	Exp	4	-0.29	0.0	I	(Normann, Emmerik, & Morina, 2014)
<b>Impairments and Psychosis</b>									
People at high risk of psychosis have impaired functioning (quality of life)	3.02	n/a	18	Quasi	2.67	-	-	I	(Fusar-Poli et al., 2015)

Patients with psychosis have elevated pro-inflammatory cytokine levels	2.21	2.45	5	Quasi	1	1.61	18.4	Q	(Upthegrove, Manzanares-Teson, & Barnes, 2014)
<b>Cerebellar peduncle (CP) in diagnosis of nervous system disorders</b>									
Apparent diffusion coefficient in the Middle CP is higher in multiple system atrophy than Parkinson's	1.66	0.97	5	Quasi	3	0.29	4.4	I	(Sako, Murakami, Izumi, & Kaji, 2016)
Superior CP is smaller in patients with progressive supranuclear palsy compared to Parkinson's	3.07	0.73	5	Quasi	3	-0.37	0.0	I	(Sako, Murakami, Izumi, & Kaji, 2017)
<b>Impaired sense of smell in Alzheimer's</b>	2.05	1.05	28	Quasi	2	0.25	2.5	I	(Rahayel, Frasnelli, & Joubert, 2012)



<b>Improving IQ scores in children</b>									
Training improves IQ but not <i>g</i>	$r = 1$	n/a	4	Quasi	2	-	-	T	(te Nijenhuis, Jongeneel-Grimen, & Kirkegaard, 2014)
Adoption improves IQ but not <i>g</i>	2.67	0.64	8	Quasi	1	-0.34	0.0	T	(te Nijenhuis, Jongeneel-Grimen, & Armstrong, 2015)
<b>Language, reading and communication skills in children</b>									
Relationship between decoding (word reading) and reading comprehension	2.20	0.20	144	Quasi	1.33	-0.64	0.0	T	(García & Cain, 2014)
Children who stutter have more negative communication attitudes than those who do not	1.83	0.83	18	Quasi	4	0.10	1.4	I	(Guttormsen, Kefalianos, & Næss, 2015)

Second language acquisition: Processing Instruction benefits grammar acquisition	2.60	n/a	42	Exp	1.67	-	-	Q	(Shintani, 2014)
L2 reading comprehension is associated with L2 vocabulary knowledge	2.58	n/a	31	Quasi	2.67	-	-	T	(E. H. Jeon & Yamashita, 2014)
<b>Visual input during tactile stimulation enhances tactile appreciation</b>	3.31	2.12	3	Exp	1	0.94	5.9	Q	(Eads, Moseley, & Hillier, 2015)
<b>Association between skill and move selection in chess</b>	2.35	0.43	4	Exp	1	-0.46	0.0	T	(Moxley & Charness, 2013)
Organisational									
<b>Relationship between relational transparency and balanced processing</b>	3.37	1.09	23	Quasi	4	-0.10	0.1	I	(Banks, McCauley, Gardner, & Guler, 2016)
<b>Association between counterproductive work</b>	2.58	0.19	46	Quasi	1	-0.76	0.0	I	(Carpenter & Berry, 2017)

behaviour and withdrawal									
Association between online IQ and consumer satisfaction	1.67	0.61	42	quasi	3	-0.07	0.3	I	(Ghasemaghaei & Hassanein, 2015)
Relationship between work engagement and burnout	1.67	n/a	25	Quasi	1.67	-	-	Q	(Maricuțoiu, Sulea, & Iancu, 2017)
Assessment of leader behavioural integrity is related to trust in leaders	2.49	n/a	22	Quasi	3.33	-	-	T	(Simons, Leroy, Collewaert, & Masschelein, 2015)
Relationship between hospitality satisfaction and loyalty	1.83	n/a	73	Quasi	2.33	-	-	I	(Tanford, 2016)
Mental Health									
Interventions for internet addition are effective	1.84	1.04	70	Exp	2	0.30	3.8	Q	(Chun, Shim, & Kim, 2017)

<b>Physical activity reduces desire to smoke</b>	2.03	n/a	17	Exp	3.67	-	-	I	(Haasova et al., 2013)
<b>Interventions for sexual violence survivors reduce clinician-assessed PTSD symptoms</b>	1.81	1.02	4	Exp	3	0.12	1.6	I	(Regehr, Alaggia, Dennis, Pitts, & Saini, 2013)
<b>Parent-Child Interaction Therapy improves externalising behaviour problems in children with disruptive behaviour disorder</b>	1.65	0.07	12	exp	2	-0.60	0.0	I	(Ward, Theule, & Cheung, 2016)
Social									
<b>Relationship between overt and relational victimisation</b>	2.08	0.48	98	Quasi	4	-0.32	0.0	I	(Casper & Card, 2017)
<b>Efficacy of group contingency interventions with children</b>	3.41	2.45	50	Exp	2	1.25	8.2	Q	(Little, Akin-Little, & O'Neill, 2015)





<b>Aerobic exercise improves executive functioning</b>	2.06	n/a	44	Quasi	0.67	-	-	Q	(Barha, Davis, Falck, Nagamatsu, & Liu-Ambrose, 2017)
--	------	-----	----	-------	------	---	---	---	---

*Note:* Quality scores run from 0-4, with higher numbers indicating higher quality. Residual heterogeneity is computed using actual minus predicted heterogeneity levels. This means that positive scores indicate excess heterogeneity (higher than expected), and negative scores indicate lower heterogeneity than expected. See Appendix C for justification of classification of effects as questionable in interest.

### 5.3.2 How meta-analyses report heterogeneity

Out of 43 meta-analyses in the very-large-effect size sample, 32 (74%) reported a measure of heterogeneity; the same proportion of meta-analyses (74%) reported a measure of heterogeneity in the control sample. In the large effects sample, heterogeneity was quantified in just over 50% of cases (23 times out of 43):  $I^2$  was reported in 21 of these cases, often alongside an additional measure of heterogeneity,  $T^2$  was reported in 7 cases, and  $T$  once. In the control sample, we assessed only whether any measure of heterogeneity was reported, and the frequency of reporting of  $T$  or  $T^2$ , and found that  $T^2$  was reported in 18 cases, and  $T$  once. Both the general reporting of heterogeneity and the quantification of heterogeneity were conveyed more frequently than in study 1a (where 55% of meta-analyses reported a measure of heterogeneity, and this was quantified in less than a third of cases). As the sample from the present study was based solely on meta-analyses that were published in the last 5 years, and the sample in study 1a was based on a broad range of dates, this could provide some tentative evidence that reporting of heterogeneity, and inclusion of a measure that quantifies heterogeneity, is improving.

### 5.3.3 Are meta-analyses of very-large effects of particularly poor quality?

As we explored earlier (see section 5.2.3), it is possible that the quality of meta-analysis conducted could affect the likelihood of a meta-analysis finding a larger effect. Higher quality meta-analyses could be more likely to take a thorough overview of the literature, to include only higher quality primary studies, to make an assessment of publication bias, and to consider the heterogeneity of studies. These factors would have the potential to have an effect on the mean effect size of a given meta-analysis: as discussed earlier, more rigorous methods are likely to introduce less sources of error, and subsequently are less likely to overestimate effect sizes such that they could be erroneously included in the very-large-effects sample. We therefore assessed whether meta-analyses in the very-large-effect size sample differed in quality from those in the control sample. Table 5.1 shows descriptive statistics for heterogeneity and quality for both the sample of meta-analyses with very-large-effect sizes, and the general/control sample. There was no significant difference in quality between the two samples,  $t(141) = 1.104$ ,  $p = .271$ . Consequently, there is not strong evidence to suggest that meta-analysis quality affects the likelihood of a meta-analysis finding a larger effect.

Quality of meta-analyses could also affect heterogeneity scores. If high quality meta-analyses are conducted with tighter inclusion criteria, and particular care is taken for accuracy in terms

of calculation of study effect sizes, this could reduce levels of unexplained heterogeneity. In support of this idea, we found a small negative correlation between  $T$  and quality for the total sample,  $r = -.264$ ,  $p = .033$  (see Table 5.3), with a similar magnitude of effect when looking at the control ( $r = -.252$ ,  $p = .103$ ) or very-large-effects ( $r = -.364$ ,  $p = .095$ ) samples separately.

#### 5.3.4 Number of studies in the meta-analysis

We considered that meta-analyses including fewer studies could result in greater uncertainty in the estimation of the true effect size, and due to this error, could be more prone to overestimating effect size. We therefore correlated  $k$  and  $d$  for the very-large-effects sample, but found no significant relationship  $r = -.13$ ,  $p = .553$  (this was also non-significant in the control sample,  $r = -.02$ ,  $p = .882$ ). This result may offer some assurance that small  $k$  meta-analyses (which inherently should tend to be of lesser quality) are not overrepresented in our very-large-effects sample.

Similarly, in study 1a, following Richard et al. (2003) we explored the idea that as a research field matures, the focus shifts from establishing an effect to exploring its boundaries. We used the number of studies ( $k$ ) as a proxy for the maturity of a research field, and correlated this with heterogeneity and found  $r_{\text{win}} = .23$  ( $p = .005$ ). We repeated this analysis in the current study, and found no significant relationship between  $k$  and  $T$  for the total sample ( $r = -.087$ ,  $p = .492$ ) or either of the samples separately (see Table 5.3).

#### 5.3.5 Meaning of meta-analyses

The majority of the meta-analyses in our very-large-effects sample appear to relate to meaningful effects (see triviality column in Table 5.4). We assessed meta-analyses investigating the relationship between highly related measures or concepts (such as intelligence) as being trivial in nature. Some meta-analyses, although investigating non-trivial effects, show high levels of heterogeneity, over and above that expected, or show that a substantial proportion of results would be expected to go against the expected direction of the meta-analytic effect. We marked these effects as questionable (see Table 5.4), and provide some illustrative examples here.

In the first example, Nishimura et al. (2013) showed that enhanced consent forms appear to improve participant understanding. However as  $FI$  for this meta-analysis shows, 24.4% of effects would be expected to go against this expected direction, suggesting that further

research may be needed before this effect could successfully be implemented in a practical setting. In an investigation of pro-inflammatory cytokine levels, Upthegrove et al. (2014) found that these are elevated in patients with psychosis, but this effect showed large excess heterogeneity (residual  $T = 1.61$ ) and a high  $FI$  (18.4%). In a perhaps particularly concerning example, Magallares & Schomerus (2015) found that bariatric surgery improves mental health related quality of life, but here excess heterogeneity was very large (residual  $T = 4.66$ ) and  $FI$  showed that the effect might be expected to reverse a high proportion of the time (11.6%). Given the physical risks associated with such surgery for patients, (for example high risk of complications, see Chang et al., 2014) understanding this effect further seems especially important.

These effects cannot be considered to be particularly consistent based on the results of these meta-analyses, and should be investigated further in order to meaningfully add to conceptual knowledge.

#### 5.3.6 Exploratory analyses on what drives heterogeneity

We have thus far been unable to determine whether the sample of meta-analyses used in study 1a might be similar to meta-analyses across the discipline, or if it might differ in some important way. We therefore thought to test for differences between the control sample in the current study and this earlier sample from study 1a. This would also allow us to assess whether similar relationships to those explored in the first study might be expected to hold in a different sample. We found no differences in mean effect size ( $d$ ) between the current study control sample ( $M = 0.43$ ) and the sample of 150 meta-analyses in study 1a ( $M = 0.54^{10}$ ;  $t(191) = 1.588$ ,  $p = .114$ ). There were also no differences between mean levels of heterogeneity for the two samples (for 150 meta-analyses in Study 1a, mean  $T = 0.34$ ; for control sample study 2, mean  $T = 0.33$ ;  $t(191) = 0.361$ ,  $p = .718$ ). It therefore appears meaningful to explore whether the relationships explored in study 1a also hold for the current study.

In study 1a, we found that mean effect size ( $d$ ) and heterogeneity ( $T$ ) correlated across multiple meta-analyses, with larger mean effects correlating with higher levels of heterogeneity. If this relationship is consistent for other sub-disciplines, we should therefore expect to find higher heterogeneity in our very-large-effect size sample than in our control sample. Consistent with this, we found significant differences between the two groups, with

---

<sup>10</sup> Note that means reported in study 1a were winsorized. Here we have reported non-winsorized means for all samples.

higher  $T$  in the sample with very-large-effect sizes ( $M = 0.79$ ) than in the control sample ( $M = 0.33$ ;  $t(77) = 4.43, p < .001$ ).

As with study 1a, heterogeneity varied widely between meta-analyses. See Table 5.2 for descriptive statistics for effect sizes and heterogeneity for the two samples. Correlating mean  $T$  and  $d$ , for all meta-analyses across both samples, resulted in  $r = .64, p < .001$ . This relationship held for both the control sample ( $r = .60, p < .001$ ), and the sample with very-large-effect sizes ( $r = .483, p = .023$ ), despite the restricted range in effect sizes (remember that all effect sizes were greater than Cohen's  $d = 1.65$ , or equivalent). In study 1a, we suggested using  $T = 0.18 + 0.30d$  to estimate heterogeneity where a meta-analysis only reports a mean effect size. Similarly to this, including all meta-analyses in the current study, the regression equation was  $T = 0.14 + 0.45d$  ( $R = .610$ ). We tested the differences between regression slopes for the control sample only, and the very-large-effect size sample only, following Paternoster, Brame, Mazerolle, & Piquero (1998) using the following formula:

$$Z = \frac{b_1 - b_2}{\sqrt{(SE_1^2 + SE_2^2)}}$$

Equation 5.1

The regression slopes calculated from the control sample only ( $b = 0.45$ ), and very-large-effect size sample only ( $b = 0.37$ ), did not differ significantly from one another ( $Z = 0.42, p = .678$ ).

#### 5.4 Discussion

Large amounts of unexplained heterogeneity have important implications for practical applications of research in psychology. High heterogeneity can make prediction of the results of a practical application (which can also be thought of as the 'next study') particularly challenging. However, larger effect sizes should mean a greater chance of success in the next study, and will minimise the risk of the effect running against the expected direction. Here we assessed the strongest effects in psychology that have been published in meta-analyses in the last 5 years. We compared these to a random, control sample of meta-analyses from the same time period in order to provide a comparison sample to determine whether different sub-disciplines show differences in the likelihood of a meta-analysis being included in our very-large-effects sample. This could be informative in terms of explaining effects (for example whether there are a few powerful causes to explain phenomena, versus many, smaller causes).

It could also be informative regarding the strength of research methods in different disciplines. The highest number of meta-analyses conducted during this period was in clinical psychology. We found a significant difference between the sample group and the number of studies in each discipline, with meta-analyses in organisational and cognitive psychology more likely to find large effects, and those in clinical and social psychology less likely to find large effects.

We discussed concerns that meta-analyses of poor quality might be systematically overrepresented in our very-large-effects sample. However, these concerns proved largely unfounded, as there were no differences in meta-analysis quality, as assessed by our adapted quality measure, between our very-large-effect size sample and control sample. However we did find a small negative association between heterogeneity and quality of meta-analyses, suggesting that higher quality meta-analyses tend to have lower levels of heterogeneity.

#### 5.4.1 Peaks of psychology – the very largest effects

The highest effect sizes related to specific treatments for mental health conditions, improvement of children's IQ and communication skills, work relationships and work behaviour, in addition to some single meta-analyses investigating other topics (see Table 5.4). A detailed discussion of each meta-analysis finding a large effect is outside the scope of this study. We therefore focus here on the areas or effects in which multiple meta-analyses showing large mean effect sizes were found in our sample.

##### 5.4.1.1 *Clinical psychology: Treatment of mental health conditions*

The inclusion of several meta-analyses in clinical psychology, and specifically several relating to the treatment of mental health conditions, is perhaps unsurprising. The contribution of mental health conditions to the global burden of disease is thought to be substantial (Lopez & Murray, 1998), and the lifetime prevalence of mental health conditions, as defined by the *DSM-IV* (Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, American Psychological Association, 1994) is thought to be around 46% (Kessler et al., 2005). There are large costs associated with this, both in terms of personal consequences for wellbeing and survival, and for society in terms of, for example, employment absences or other employment struggles such as loss of productivity (for example, see Goetzel et al., 2004). In light of this, identification of the most efficacious treatments for mental health conditions has been identified as a top research priority in mental health (Collins et al., 2011).

More specifically, as relatively new treatments, the effectiveness of repetitive transcranial magnetic stimulation (rTMS) and deep transcranial magnetic stimulation (DTMS) are still being explored for conditions such as persistent PTSD and depression, and the meta-analyses in these topics could reflect this. However, the authors themselves note that the findings are preliminary in nature, and outcomes are measured a short time after treatment – there is no assessment of the longer term effectiveness of these techniques at present. Meta-analyses investigating the most effective treatments for depression and anxiety are also perhaps unsurprising, and specific therapies such as cognitive behavioural therapy (CBT) have received a great deal of attention, having been found to be particularly cost effective, in financial terms, for a range of conditions (for example, see Mccrone et al., 2004; Vos, Corry, Haby, Carter, & Andrews, 2005). It is perhaps interesting to note that the very-large-effects seen in the meta-analyses from our very-large-effects sample investigating these interventions in relation to ratings of symptoms often relate to clinician-rated symptoms (Jonsson et al., 2015; Rosa-Alcázar et al., 2015). On our examination of moderator analyses reported within these meta-analyses on self-reported symptoms, it is apparent that these do not always translate to the same magnitude of symptom relief as experienced by individuals with a given condition (the effect size for self-reported symptom ratings were half that of clinician ratings). Had the meta-analyses been conducted on individuals' own ratings of their symptoms, the effects would not have been large enough to be included the current sample. Some meta-analyses do not differentiate between these forms of symptom rating, which could further confound this issue (for example, see Bandelow et al., 2015). This may be a particular issue if researchers conducting the original studies, and clinicians administering treatment, are not blinded as to allocation of participants to treatment conditions, which could introduce bias in recording of symptom severity. The included meta-analyses do not make an assessment of the possible biases in the original studies.

Bias relating to symptom reporting is not an issue in studies utilising more objective measures, for example measurement of water diffusion in the cerebellar peduncle in parkinsonian disorders (Sako et al., 2016, 2017). While meta-analyses do not assess possible biases, it is not possible to determine how much of an influence bias could have on meta-analytic results; however, use of more objective measures does appear to make these effects more objectively promising.



#### 5.4.1.2 *Cognitive psychology: Improving IQ and language, reading and communication skills in children*

In cognitive psychology, several meta-analyses considered the improvement of IQ scores in children, as well as children's language, reading and communication skills. Both meta-analyses looking at IQ scores in children were correlational in nature, and considered whether improvements in intelligence (either due to adoption into a more prosperous family, or due to training designed to increase academic achievement) were related to increased general intelligence (*g*) as measured using tests considered to have high *g*-loadings (sometimes conceptualised as having higher cognitive complexity, for example see te Nijenhuis et al., 2015). Both meta-analyses have the same first two authors, and note that many tests of IQ share common variance and are often highly correlate; this may explain the high effect sizes found in these meta-analyses.

In terms of the meta-analyses looking at children's skills, these sought to understand key factors in children's learning. Meta-analyses examined the relationship between word reading and reading comprehension, grammar knowledge and reading comprehension (in either English or a second language), the importance of the type of instruction to learning in a second language, and barriers to communication (specifically relating to attitudes to communication). More specifically, Guttormsen, Kefalianos, & Næss (2015) found large differences in attitudes toward communication in children who stutter and those who do not, but also noted large heterogeneity between the studies.

In terms of the associations examined in the meta-analyses on reading and language, the processes involved in, for example, word reading, vocabulary knowledge, and comprehension, are closely related (for example, see Jenkins, Fuchs, Van Den Broek, Espin, & Deno, 2003; Walczyk, 2000), and the strength of the relationships between these constructs are perhaps therefore unsurprising. In their meta-analysis on second language grammar acquisition, Shintani (2014) notes that despite the apparently large effect of their intervention, the possible reasons for this are still poorly understood, but could relate to close relationship between the type of practice being used in the intervention (comprehension) and the skills being assessed by the specific test which achieved such a large effect, as the results don't generalise to other outcome measures.

#### 5.4.2 Error in meta-analyses

The mean effect size obtained via meta-analysis will inevitably reflect both the true effect size underpinning the relevant phenomenon, and a certain amount of error. The amount of error included will relate to the effect sizes reported by the primary studies included in the meta-analysis. As discussed previously, because we selected the largest effects in meta-analyses in psychology in the last five years, we should expect that the magnitude of these effects will tend to be an overestimation of the true effect size due to regression to the mean. This is not an issue for the aim of our study, as observed very-large-effects remain our most promising avenue of exploration when trying to locate the largest true effects. The inclusion of multiple meta-analyses into the same effect in our very-large-effects sample could provide some assurance that these effects are not down to error in a specific meta-analysis.

#### 5.4.3 Distribution of large effects across sub-disciplines

The discrepancy in the proportion of meta-analyses in clinical psychology between the very-large-effect size sample and the control sample could be due to problems with the cut-off criteria we used for inclusion of very-large-effect sizes. This was based on samples from study 1a, which included the sub-disciplines of cognitive, organisational and social psychology. While effect size measures based on standardised mean difference (e.g. Cohen's  $d$ ) and correlations are particularly common in these sub-disciplines, OR and Log OR are used more frequently in disciplines such as clinical psychology which is perhaps particularly strongly influenced by medical disciplines that utilise similar effect size measures (Ferguson, 2009; Kazis, Anderson, & Meenan, 1989; Rutledge & Loh, 2004). While the conversion of effect sizes from one metric to another is mathematically possible, it is perhaps not conceptually sensible to do so due to issues such as differences between continuous and dichotomous outcomes, and that this could result in an effect size conversion that is not particularly realistic for these types of effect sizes (Ferguson, 2009). It is therefore important to interpret this aspect of our findings with caution. For the other sub-disciplines, which more commonly utilise effect size measures such as Cohen's  $d$  or Pearson's  $r$ , social psychology was less frequently represented in the sample of meta-analyses with very-large-effects (7% of this sample, compared to 13% of the control sample), whereas cognitive and organisational psychology were more frequently represented (19% versus 2%, and 14% versus 7%, respectively).

In addition to this, there could be differences in the mean sample sizes of studies included in meta-analyses from different sub-disciplines. As sample size and effect sizes of studies are

often thought to be linked (Bakker, van Dijk, & Wicherts, 2012; Ferguson & Heene, 2012), this could, in turn, affect the effect sizes found in meta-analyses (Fanelli & Ioannidis, 2013), perhaps by over-inflating effect sizes in some instances (when meta-analyses are primarily based on smaller studies). This could not be addressed here, however instead we considered the possibility that meta-analyses with fewer studies might provide less accurate estimations of the true effect size, and might be more likely to overestimate true effect size due to including greater error, but, reassuringly, found no relationship between  $d$  and  $k$ .

#### 5.4.4 Comparisons with the sample from study 1a

We also found similarities between the control sample of meta-analyses in the current study and the sample we utilised for our overview of heterogeneity in study 1a, in terms of both the mean effect size, and mean level of heterogeneity. A key relationship explored in study 1a also appeared to hold, namely the positive relationship between effect size and heterogeneity, which also led to similar predictions of heterogeneity based on effect size. However, the relationship between the number of studies in a meta-analysis (used as a proxy for maturity of research field) and heterogeneity that was found in Study 1a, did not hold in the current study, and in fact was reversed in direction. This parallels the findings from study 1a, which indicated that the reasons for this apparent relationship in this, and a previous, sample (see Richard et al., 2003) are not well understood.

#### 5.4.5 Limitations

The main limitation for this study was the fact that meta-analyses utilising certain effect sizes which are commonly used in certain sub-disciplines of psychology (namely OR and similar) were not included in the sample of studies with very-large-effects, as no studies utilising these types of effect sizes were greater than our selected effect-size cut off point. This could mean that the large effects identified in this study are not truly representative of the largest effects demonstrated across psychology as a whole, and instead only represent the largest effects presented as either Standardised Mean Differences (Cohen's  $d$  or Hedges  $g$ ) or correlations (Pearson's  $r$  or Fishers  $Z$ ).

In addition to this, this study was only able to consider phenomena for which meta-analyses have been conducted. Where strong effects are described, but data are not presented to allow effect sizes to be calculated, there will be no meta-analysis to describe such effects. For example, operant conditioning appears to be a very powerful method of shaping behaviour in

pigeons (Skinner, 1948), however, results of these experiments are presented in a descriptive manner, and the results appear to be sufficiently strong that researchers in this field might not see the interest in conducting a meta-analysis on operant conditioning in pigeons.

Furthermore, for feasibility reasons, only meta-analyses from the last 5 years were included in this study. While new meta-analyses into specific phenomena are being conducted all the time, there could be effects that are relatively well established with well-conducted meta-analyses from an earlier time point that have not been repeated within this time frame. This means that our sample may not be representative of psychology as a discipline, as this spans a far broader time period, and our sample does not include effects that might have been well established more than 5 years ago. It is highly unlikely that very-large-effects have only been meta-analysed within the last 5 years, and so our analysis provides only a snapshot of recently meta-analysed very-large-effects.

For our heterogeneity analyses, we were only able to consider meta-analyses for which we could obtain sufficient detail regarding the effects in underlying primary studies, which greatly reduced our sample sizes. This also relies on reported effect sizes for the underlying primary studies. As with Study 1a, it is not clear whether the remaining sample is representative of all meta-analyses, or how much error reliance on reported effects could introduce.

#### 5.4.6 Conclusion and future directions

Our results show that high numbers of meta-analyses are conducted in clinical psychology, and that meta-analyses in organisational and cognitive psychology are more likely to find large effects, while those in social psychology are less likely to find large effects. Further investigation of heterogeneity in relation to other effect sizes, such as Odds Ratios, could usefully be conducted to establish whether similar principles can apply to meta-analyses utilising these measures, and to assess whether our conclusions can generalise to other sub-disciplines in which such effect sizes are common.

Another initial key observation is that focus on effect size alone can provide a skewed perspective on what we know about a given phenomenon. There are some areas where there appear to be consistency in the direction of effects, but mean levels of heterogeneity in our very-large-effect sample were, unsurprisingly, very high. Although mean  $T$  was 0.79,  $T$  ranged from 0 to 7.54. At this higher end of heterogeneity, it is arguable how well an effect is understood, unless the sources of such heterogeneity can be identified. True effects in this

case, even with a mean  $d$  of 2.28, could run from a medium effect to a very-large-effect. As discussed previously, very-large-effects several effects showed excess heterogeneity over and above that expected, and some showed that a proportion of results would be expected to run counter to the expected direction of this effect. As a counterpoint, it is worth noting that many of the included meta-analyses showed lower heterogeneity than might be expected, suggesting that these results might be particularly consistent.

No differences in quality were found for our large effects versus our control sample, but higher quality meta-analyses do seem to have lower levels of heterogeneity, suggesting the importance of following key guidelines for conducting meta-analyses well (for example, following the PRISMA guidelines closely, Moher et al., 2009) for reducing levels of observed heterogeneity. We found that heterogeneity does seem to be being quantified more frequently than in the meta-analyses from study 1a, which may suggest that the value of this is being increasingly recognised.

We also found that similar effect sizes and mean levels of heterogeneity were present in our control sample of meta-analyses which utilised similar effect sizes to those in the sample from study 1a. This could provide some indication that these mean effect sizes and heterogeneity levels can be generalised more widely, and that the prediction of heterogeneity level based on effect sizes could also be more widely applicable.

The current study took a relatively novel approach by examining the largest effects in psychology. There is a commonly held belief that worthwhile research must necessarily be theory-driven (Greenwald, 2012), and here we took an alternative, descriptive approach. We therefore find it valuable to reflect on the usefulness of this perspective and what it can add to existing approaches.

Similar work cataloguing observations in this way is considered key in other research fields, for example, after other chemists had begun to record the properties of chemical elements, Mendeleev was able to recognise that arranging and categorising them using their relative atomic weights created a meaningful periodic table in chemistry (Pfennig, 2015). This can also allow patterns in empirical data to be observed: to continue our example, the periodic table is arranged by atomic number, and shows patterns in terms of families of elements that group together based on similar chemical properties (for example, the noble gases, which are grouped into a single column in the periodic table, are all found in air, and are all very unreactive and colourless, Pfennig, 2015). When elements were still being identified and

catalogued, this elucidated areas where elements appeared to be 'missing' from the configuration, and their subsequent discovery/identification confirmed theoretical propositions regarding these elements' existence and their properties. Bringing together multiple meta-analyses in psychology could therefore also begin to illuminate patterns or raise questions for subsequent investigation (for example, we noted earlier that several meta-analyses appear to find similar magnitudes of effects for similar phenomena). Work that paved the way for Mendeleev's creation of the periodic table was valuable for this reason, yet the value of recording the properties of elements was not anticipated until Mendeleev was able to put this work to use. Similarly, the work undertaken for the current study is only a starting point. Other researchers will need to determine whether this work has sufficiently piqued their curiosity to continue in this direction.

## 6 Study 3: Heterogeneity of sex differences across cultures

### 6.1 Introduction

Study 2 considered the applicability of heterogeneity to the issue of the effectiveness and likely success of practical applications of research, however, heterogeneity can also be utilised in other contexts. One domain in which heterogeneity could prove useful is to provide a novel source of evidence to evaluate the origins of sex-differences in psychology.

Sex differences in psychology have been observed in relation to a variety of traits, from personality traits such as dominance (men tend to be more dominant than women,  $d = 0.26$ , Feingold, 1994) or empathy (women tend to show more empathy than men,  $d = 0.97$ , Feingold, 1994) to cognitive abilities such as navigational ability (men tend to perform better than women,  $d = 0.29$ , Coutrot et al., 2018) or mental rotation (men tend to perform better than women,  $d = 0.47$ , Richard A Lippa, Collaer, & Peters, 2010), mathematics achievement and ability (estimates of sex differences differ across studies, with boys sometimes outperforming girls and vice versa, for example see Else-Quest et al., 2010; Lynn & Irwing, 2008), mate preference (e.g. males value good looks in a mate more than females do, see for example Buss, 1989) and even prevalence of specific diagnoses, such as autism spectrum condition (men have higher rates of autism diagnosis than women,  $OR = 4.2$ , Loomes, Hull, & Mandy, 2017). However, the underlying causes for these differences are strongly contested (for example, see Eagly & Wood, 1999; Schmitt, 2015). Principally, there are two main perspectives, the evolutionary perspective and the social structuralist perspective.

#### 6.1.1 Evolutionary perspective

The evolutionary perspective asserts that at least some sex differences are evolved and hard-wired, including some psychological sex differences. Evolutionary psychologists suggest that prehistorically, women and men faced different challenges regarding survival and reproduction, and that gender was key to these adaptive challenges (Schmitt, 2015). Differing social roles occupied by men and women then stem from sex-specific adaptations that lead men and women to differ psychologically. This viewpoint asserts that numerous sex differences have been empirically observed that are of meaningful (moderate to very large) effect size, and that these are consistent across different cultures (Archer, 2019; Schmitt, 2015). For example, males value good looks in potential mates more than females do, and males prefer mates who are younger (see Table 6.1 for both of these effects), and these

effects are both consistent across cultures (all effect sizes are in the same direction). We refer to this perspective as the consistency argument.

A variation on this perspective suggests that evolved traits could differ meaningfully across cultures due to differences in what would be adaptive depending on specific regions or cultures (Schmitt, 2015). For example, it has been argued that facial symmetry can be an important aspect of mate choice, as this could be an outward indication of having a strong immune system (Scheib, Gangestad, & Thornhill, 1999). This preference might be particularly adaptive in areas of high pathogen load. In support of this, stronger relationships have been found between facial symmetry and attractiveness of opposite-sex faces when people are exposed to cues relating to environmental pathogens (Little, DeBruine, & Jones, 2010; Little, DeBruine, & Jones, 2013). A similar perspective might hold for evolved sex differences.

#### 6.1.2 Social structuralist perspective

In contrast to this, the social structuralist perspective suggests that observed psychological sex differences arise due to differences in social roles, leading to specific patterns of preferences (Eagly & Wood, 1999). Situations faced by women and men vary quite extensively. Male and female roles have also differed over time within countries, and can differ between different cultures, and these differences can drive observed sex differences. Sex differences are then a tangible indication of the opportunities and restrictions placed on women and men due to their differing social roles. For example, there is a sex difference in mean number of hours worked (women work less hours on average than men), and in part this could stem from societal expectations that women should undertake more responsibilities within the home, and that men are responsible for financially supporting their family (MacInnes, 2005). From this viewpoint, we might expect that the magnitude of sex differences will vary across different cultures with different social characteristics, and that this will be reflected in a strong correlation between magnitude of sex difference and a cross-cultural factor (such as gender equality). We refer to this perspective as the covariation argument.

In line with the social structuralist perspective, Hyde (2005) has proposed the gender similarities hypothesis, which asserts that males and females are similar on most psychological variables, and that gender differences can fluctuate in size depending on variables such as the context of measurement or participant's ages. Examining multiple meta-analyses of sex differences, Hyde illustrates that 78% of sex differences are small or close to zero, leaving few sex differences with moderate to large effect sizes: of these, it is arguable that domains such



as motor performance are strongly influenced by physical sex differences. In support of this, in a more recent large meta-synthesis of 106 meta-analyses of sex differences, Zell et al. (2015) argue that in many psychological domains, men and women are more similar than they are different, and that the average sex difference is small ( $d = 0.21$ ). Zell et al. (2015) argue that perceptions of sex differences are larger than the reality of these differences due to cultural influences around gender stereotypes, and the conflation of physical differences with psychological differences. In contrast to this, they argue that actual sex differences are due to the influence of societal forces.

### 6.1.3 Why both the consistency argument and the covariation argument are flawed.

To better understand each perspective, it is first important to understand the concept of the environment of evolutionary adaptedness (EEA). The EEA refers to an ancestral environment over 10,000 years ago that proponents believe had remained relatively stable for 1 to 2 million years previously (Bolhuis, Brown, Richardson, & Laland, 2011; Irons, 1998). The evolutionary perspective suggests that human adaptations evolved in response to features of the EEA. These adaptations would have conferred an adaptive advantage in the EEA, and in environments which are similar to the EEA, but would not necessarily confer the same advantage in current environments (Irons, 1998). Proponents of the EEA posit that this environment would have led to different pressures on males and females, both physically and psychologically, and that these pressures led to evolved sex differences.

Returning to the two perspectives introduced above, remember that the consistency argument (put forward in favour of the evolutionary perspective) proposes that consistency of effect sizes (all, or most, effect sizes for sex differences are in the same direction) across cultures suggests that these sex differences are caused by evolutionary processes. However, we believe that this argument is insufficient to support the idea that sex differences are caused by evolutionary processes. Consider the fictitious scenario in the left panel in Figure 6.1. This shows a strong mean sex difference in a given attribute (for example, men prefer mates who are younger than them), with little variation across current cultures. Current societies differ from the EEA. A crucial question is therefore, how different would the EEA have to be from present-day environments on the cultural factor for this sex difference to disappear? If we consider the regression line for this sex-difference, we can see that the EEA would have to be very greatly different from present-day environments before the sex difference reached the intercept with the x-axis (no sex difference). Compare this to another fictitious scenario showing a similarly large sex difference across current cultures, but this time with a large

degree of variability across cultures (right hand panel in Figure 6.1). In this case, we can see that the EEA would not need to be particularly different from today's environments in order for the sex difference to disappear. In other words, although the consistency argument holds for both scenarios, only the left provides sufficient support for a sex difference in the EEA. The crucial difference between both scenarios is, of course, that cross-cultural heterogeneity is low for the left one but high for the right one.

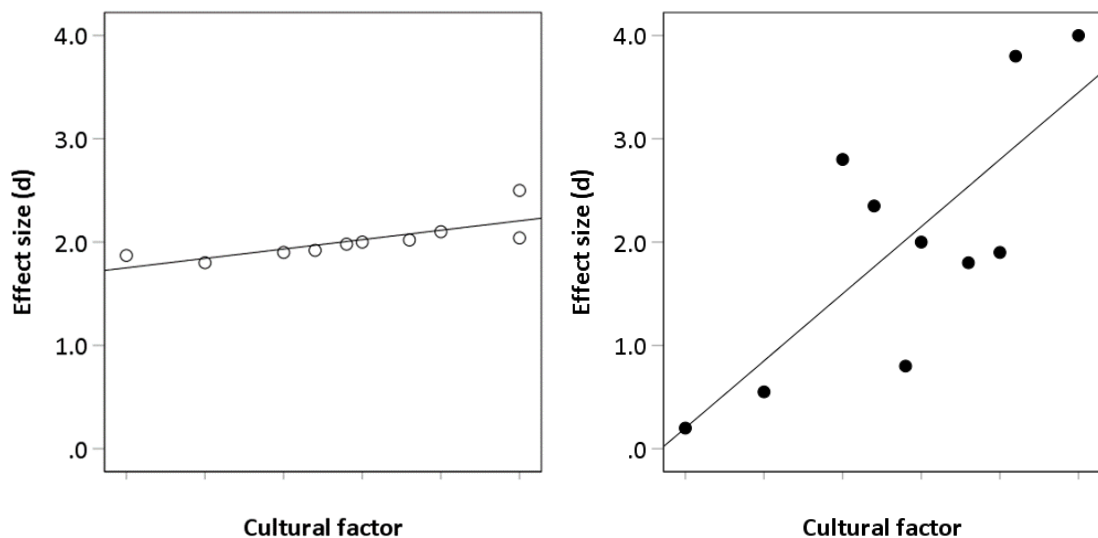


Figure 6.1 Two fictitious patterns of a sex difference across cultures. *Left hand panel*: large mean effect size with low variability in effects ( $M = 2.01$ ,  $SD = 0.19$ ); here, a sex difference for the EEA seems more plausible because the EEA would have to be very greatly different from present-day environments for the sex difference to disappear. *Right hand panel*: large mean effect size with high variability in effects ( $M = 2.02$ ,  $SD = 1.28$ ); here, a sex difference for the EEA seems less plausible because the EEA would not have to greatly differ from present-day environments for the sex difference to disappear.

From a social-structuralist perspective, the covariation argument suggests that sex differences which show cross-cultural covariation with a particular characteristic of society (for example, sex differences that change in magnitude in accordance with differences in gender equality) are sufficiently explained by this characteristic. Similarly to the consistency argument, we believe that the covariation argument is not very strong. Consider Figure 6.1 left panel once more. This shows a strong correlation between the magnitude of the sex difference and a crucial cultural factor ( $r = .77$ ). However, this correlation is equally strong in the right panel ( $r = .77$ ). Covariation between sex differences and cultural differences is therefore insufficient to support the social structuralist perspective.

#### 6.1.4 Heterogeneity of sex differences

The shortcomings of both the consistency and covariation arguments that we have outlined above illustrate that variability of the effect sizes across cultures is a key piece of data that should be considered to understand the likely origins of sex differences. This variability can be assessed by determining the level of heterogeneity that applies to a given sex difference across cultures. As this application of heterogeneity has yet to be examined, this study aimed to explore whether heterogeneity could usefully be applied to provide a new perspective on the issue of the underlying cause of sex-differences. We aimed to examine a number of psychological sex differences where data are available from different nations and cultures.

To properly evaluate these data, it is helpful to have a comparison standard. Therefore, data regarding height (e.g. Eveleth, 1975; Gray & Wolfe, 1980; Gustafsson & Lindenfors, 2009) and 2d:4d digit ratio (indicative of prenatal testosterone levels; e.g. Hönekopp & Watson, 2011), were used to represent biological sex differences. 2d:4d digit ratio (the ratio of the length of second to fourth fingers) is used as an indirect measure of prenatal testosterone levels, as 2d:4d is thought to show a negative relationship with prenatal testosterone, and this relationship is stable during development (Hönekopp & Watson, 2010). This means that males generally have lower 2d:4d than females. Both sex differences in height, and those in prenatal testosterone, are thought to be consistent with evolved and adaptive responses to prehistorical environments (Gaulin & Boster, 1985; Richard A. Lippa, 2009), and we are not aware of any discussions of these sex differences that treat them as having social origins. Height and 2d:4d digit ratio were therefore used to determine the levels of heterogeneity for a sex differences that are widely thought to be biological in nature.

As a second form of comparison standard, for sex differences widely thought to be cultural in nature, we used data from the Programme for International Student Assessment (PISA; for a similar example, see Mann, Sasanuma, Sakuma, & Masaki, 1990) relating to students' interest in science and careers in science, technology, engineering and mathematics (STEM), and data regarding adults' attitudes toward gender role and behaviours based on the International Social Survey Program (ISSP; see Stickney & Konrad, 2007). Equality of gender opportunities and attitudes are widely used as markers for the cultural influences on sex differences (for some examples, see Batz-Barbarich, Tay, Kuykendall, & Cheung, 2018; Else-Quest et al., 2010; Schwartz & Rubel-Lifschitz, 2009; Stoet, Bailey, Moore, & Geary, 2016), and we are not aware of any discussion of these differences that treats them as having biological origins. We

therefore used these data to provide a comparison standard for sex differences that are thought to be primarily cultural in nature.

We aimed to use the data outlined above as a comparison standard for levels of heterogeneity that can be expected from primarily biological or primarily cultural differences, in order to inform our analysis of data on psychological sex differences. Our expectation was that we would see greater levels of heterogeneity across cultures in sex difference that were primarily cultural in nature (see Figure 6.2). In contrast to this, we expected to see lower levels of heterogeneity in sex differences that are widely thought to be biological in origin and not greatly influenced by cultural factors (see Figure 6.2).

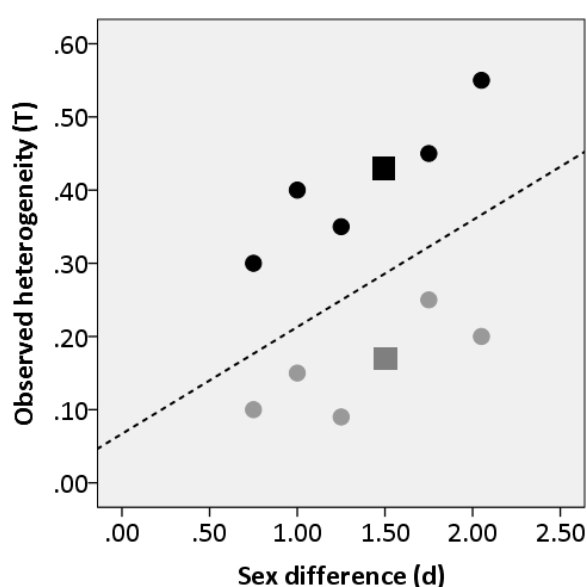


Figure 6.2 Expected pattern of sex differences (based on fictitious data). Biological sex differences (grey circles), and cultural sex differences (black circles). Square symbols represent sex differences from our psychological sex difference sample. The grey square represents a psychological sex difference for which we would infer biological origins; the black square represents a psychological sex difference for which we would infer cultural origins.

## 6.2 Method

### 6.2.1 Study search and selection strategy

We intended to search for psychological sex differences where data were available on the same measure from at least 8 different countries. For example, the same questionnaire, assessment tool, or identical methodology were used to collect the data from samples in

different countries. This should help to avoid confusing real heterogeneity in sex differences with artificial sources of heterogeneity arising from differences in methodological approaches.

We searched PsycINFO, journals only, and WebOfScience for (“meta-analy\*” OR “cross cultur\*” OR “cross nation\*” OR “global variat\*”) AND (“sex\* differ\*” OR “sex\* dimorph\*” OR “gender similar\*” OR “gender differ\*”), in September 2018. This resulted in 2469 search results (see Figure 6.3).

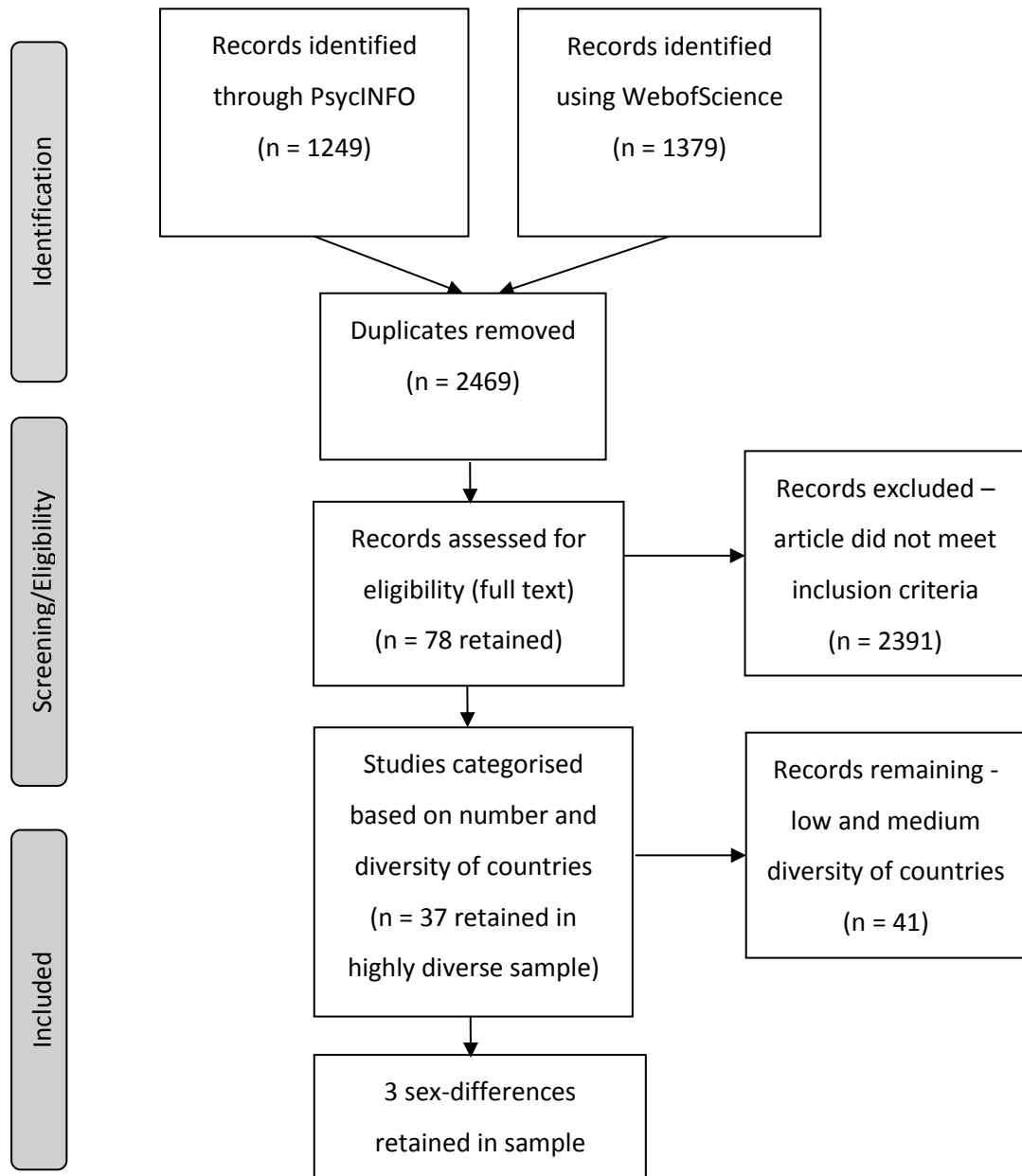


Figure 6.3 Search and selection of effects for sex differences

#### 6.2.1.1 *Inclusion criteria*

Results were screened (full text) and included in the next stage if they met all of the following criteria: i) study must include measurement of a sex difference, and it must be possible to calculate a standardised mean difference for this sex difference. ii) Data must be included from at least 8 countries, and this data must be provided for each country separately. iii) All studies in the paper must use the same protocol. iv) For practical reasons, at this stage the full article had to be available in English.

After this stage of screening, 78 articles were retained (see Figure 6.3). These were then filtered into categories based on the number and diversity of countries included in each paper. Diversity of countries was assessed based on geographical region, for example inclusion of data from both western and non-western countries, and from different continents. We retained only those articles which fell into the highly diverse sample, leaving us with 37 articles. From this we selected 3 domains: mate preferences (5 sex differences), mental rotation (2 sex differences), and mathematics ability (3 sex differences). This selection was made because these sex differences are widely discussed (for example, see overviews by Archer, 2019; Schmitt, 2015), yet have ambiguous status in terms of their social or biological underpinnings, and for reasons of feasibility and data availability.

#### 6.2.2 Data sources for sex differences

For each of our 15 datasets for sex differences (10 psychological, 2 biological, and 3 cultural) we retained only one sample per country in our analysis. For most of the included sex differences (all sex differences relating to mate preference, mental rotation, mathematics ability, interest in science, future career, and gender attitudes), data were collected by the original authors of the study as part of a single study or survey and presented on a country-by-country basis. To retain this approach, we included only the largest sample available for each country for our other sex difference examples as well (height and 2d:4d). For each sex difference we extracted the mean and standard deviation for each effect, as well as the sample size for each group from the meta-analyses for each country, and the same information for any primary studies included. In some cases standard deviations had to be computed from reported standard errors (see section 2.1.2).

### 6.2.3 Data for psychological sex differences

#### 6.2.3.1 *Sex differences in mate preferences*

All data for calculation of these sex differences were based on a large cross-national study and meta-analysis, including samples from 37 countries, conducted by Buss et al.(1989).

Researchers in each country were instructed to recruit a diverse range of participants, however, Buss et al. (1989) acknowledge that rural, less educated, and lower socio-economic status participants are underrepresented. The questionnaire examined multiple aspects of mate preference using the same questionnaire, translated into multiple languages.

Participants were asked to rate a series of 18 characteristics on how important or desirable it would be when choosing a mate. For example, the importance of their spouse having good looks. Items were assessed on a 4-point scale from 3 (indispensable) to 0 (irrelevant/unimportant). Mate preferences included in the current analysis are the target sex differences reported in Buss et al. (1989), namely mate preferences in terms of: good financial prospect; ambition and industriousness; age difference between self and spouse; good looks; and chastity (see Table 6.1). For age difference between self and spouse, participants were asked to provide the ideal age gap that they would prefer between themselves and their spouse, as well as who they would prefer to be older (self or spouse).

Table 6.1 Sample sizes and effect sizes (*d*) for sex differences on five aspects of mate preference

<b>Country/ population</b>	<b>Total N</b>	<b>Good financial prospect</b>	<b>Ambition and industrious- ness</b>	<b>Age difference between self and spouse</b>	<b>Good looks</b>	<b>Chastity</b>
Australia	280	1.09	0.54	1.63	-0.56	-0.32
Belgium	145	0.47	0.35	2.00	-0.62	-0.34
Brazil	630	0.81	0.60	2.00	-0.26	-0.62
Bulgaria	269	0.52	0.56	1.93	-0.57	-0.28
Canada (English)	101	1.12	0.72	2.44	-0.53	-0.28
Canada (French)	105	0.67	0.37	2.10	-0.42	-0.37
China	500	0.48	0.55	0.92	-0.72	0.09
Colombia	139	0.60	-0.14	2.38	-0.44	-1.16
Estonian S.S.R.	303	0.23	0.23	4.37	-0.92	-0.41



Finland	204	0.65	0.16	2.38	-0.76	0.03
France	191	0.49	0.26	2.60	-0.40	-0.05
Germany-West	1083	0.74	0.31	1.85	-0.82	-0.27
Great Britain	130	0.67	0.53	1.88	-0.88	0.03
Greece	132	0.87	0.32	2.89	-0.38	-0.09
India	247	0.49	0.81	0.94	-0.08	-0.26
Indonesia	143	1.47	0.46	2.70	-0.61	-0.07
Iran	55	0.83	0.25	3.69	-0.54	-0.47
Ireland	122	0.99	0.38	1.42	-0.97	-0.02
Israel (Jewish)	473	0.55	0.77	1.82	-0.25	-0.34
Israel (Palestinian)	109	0.40	0.41	2.37	-1.27	-1.12
Italy	101	0.61	0.49	2.13	-0.47	-0.52
Japan	259	2.09	0.68	2.54	-0.55	-0.67
Netherlands	417	0.30	0.14	1.00	-0.76	0.00
New Zealand	151	0.32	0.44	1.43	-0.99	-0.15
Nigeria	172	1.16	0.56	1.97	-0.61	-0.71
Norway	134	0.35	0.12	1.76	-0.66	-0.01
Poland	240	0.80	0.46	1.81	-0.20	-0.23
S. Africa (whites)	128	1.01	0.57	0.96	-0.55	-0.20
S. Africa (Zulu)	100	0.53	-0.40	0.52	-0.39	-0.98
Spain	124	0.15	-0.04	1.10	-0.87	-0.37
Sweden	172	0.69	0.09	2.08	-0.24	0.05
Taiwan	566	1.27	0.95	2.99	-0.67	-0.14
USA (Hawaii)	179	0.80	0.42	1.45	-0.72	-0.37
USA (Mainland)	1491	1.04	0.90	1.63	-0.64	-0.37
Venezuela	193	0.69	0.29	1.78	-0.52	-0.33
Yugoslavia	140	0.52	0.57	3.71	-0.66	-0.63
Zambia	119	1.09	0.20	1.49	-0.69	-0.66

*Note:* Countries/populations are as reported in Buss (1989). Positive values for Cohen's *d* reflect higher scores in females.

#### 6.2.3.2 *Sex differences in mental rotation and line angle judgment*

The analysis of this sex difference was based on a large cross-national study examining mental rotation and line angle judgements, undertaken by the BBC in an internet survey in 2005 (Lippa, Collaer, & Peters (2010). The relevant data were not available in the publication but

were shared by Lippa upon request. Participants completed a 6-item mental rotation test (Peters et al., 1995), and a 20-item line angle judgement test (adapted from Benton, Varney, & deS Hamsher, 1978; Collaer, 2001). Both the mental rotation test and line angle judgement test were presented online. For the mental rotation test, participants were presented with an image of an object, and asked to judge which two of four diagrams next to this showed the same object, but “viewed from a different angle”. All six items had to be completed within a single 150 second time limit. A single point was awarded for each correct answer, creating a total score ranging from 0 to 12. For the line angle judgement test, participants were presented with a target line, displayed in one of 14 angled orientations at approximately 12.9 degree increments from horizontal to vertical. A diagram showing all possible line orientations was displayed below, and participants were given 10 seconds to select from this the line that matched the target line in orientation. Scores were the number of correct answers, with possible scores ranging from 0-20. We included adult samples from the 53 countries sampled in this paper in our analysis of line angle judgement, and 52 countries in the analysis of mental rotation (see Table 6.2).

Table 6.2 Sample sizes and effect sizes (*d*) for sex differences in mental rotation and line angle judgment tasks

Country/ population	Line angle judgement		Mental rotation	
	Total N	<i>d</i>	Total N	<i>d</i>
Argentina	226	-0.64	264	-0.47
Australia	8016	-0.58	7765	-0.46
Austria	400	-0.32	372	-0.40
Belgium	1319	-0.40	1474	-0.44
Brazil	288	-0.35	270	-0.49
Bulgaria	378	-0.55	360	-0.66
Canada	11718	-0.57	13165	-0.46
Chile	104	-0.54	124	-0.52
China	302	-0.39	400	-0.73
Croatia	210	-0.75	183	-0.06
Cyprus	114	-0.60	113	-0.84
Czech Rep	253	-0.66	279	-0.54
Denmark	780	-0.48	867	-0.48
Egypt	154	-0.45	150	-0.20

Estonia	141	-0.47	138	-0.08
Finland	1629	-0.52	1822	-0.37
France	978	-0.40	1060	-0.61
Germany	1508	-0.44	1628	-0.39
Greece	844	-0.55	741	-0.49
Hungary	204	-0.34	221	-0.49
Iceland	178	-0.70	213	-0.50
India	3213	-0.46	3219	-0.28
Ireland	5216	-0.74	5480	-0.48
Israel	358	-0.75	405	-0.43
Italy	454	-0.57	470	-0.65
Japan	481	-0.55	500	-0.43
Lithuania	125	-0.17	142	-0.29
Malaysia	771	-0.38	627	-0.53
Malta	107	-0.62	135	-0.61
Mexico	344	-0.44	393	-0.41
Netherlands	2126	-0.50	2021	-0.49
New Zealand	2002	-0.50	1664	-0.48
Norway	565	-0.60	773	-0.50
Pakistan	351	-0.41	371	-0.04
Philippines	413	-0.39	423	-0.33
Poland	446	-0.54	486	-0.45
Portugal	343	-0.79	389	-0.51
Romania	364	-0.62	384	-0.48
Russian Federation	189	-0.31	209	-0.24
Saudi Arabia	87	-0.24	100	-0.38
Singapore	1787	-0.38	2319	-0.51
Slovenia	258	-0.33		
South Africa	84	-0.07	370	-0.47
Spain	820	-0.49	868	-0.56
Sweden	1246	-0.66	1381	-0.58
Switzerland	557	-0.52	587	-0.42
Thailand	150	-0.16	163	-0.37
Trinidad and Tobago	132	-0.40	143	-0.42
Turkey	1267	-0.45	1351	-0.41

United Arab Emirates	304	-0.71	322	-0.41
United Kingdom	96044	-0.60	91837	-0.51
USA	61295	-0.49	95745	-0.48
Venezuela	89	-0.43	130	-0.30

*Note:* Countries/populations are as reported in Lippa (2010). Negative Cohen's *d* means that men have higher scores.

### 6.2.3.3 *Sex differences in maths ability*

#### 6.2.3.3.1 PISA 2015 – age 15

We used the most recent available data from the Programme for International Student Assessment (PISA), 2015 (*PISA 2015 Database*, 2015). This worldwide survey assesses the scholastic performance of 15 year olds across the globe in relation to mathematics, science, and reading performance, and is undertaken by the Organisation for Economic Co-operation and Development (OECD) every 3 years. Data from this survey are publicly available for download on the PISA website. We used data for each of 69 participating countries, by gender, for the PISA mathematics scale (overall score), which shows good reliability (Cronbach's  $\alpha = 0.85$ , Organisation for Economic Co-operation and Development, 2017). See Table 6.3 for samples and effect sizes.

#### 6.2.3.3.2 TIMSS 2015 – 4<sup>th</sup> grade

We used the most recent available data from the Trends in International Mathematics and Science Study (TIMSS), 2015 (*TIMSS 2015 International Database*, 2015). This worldwide survey evaluates mathematics and science knowledge in a minimum of 4,500 students in each participating educational system, and is administered every 4 years by the International Association for the Evaluation of Educational Achievement (IEA). Data from this survey are publicly available for download from the TIMSS website. We used data for each of 48 participating countries, by gender, for the mathematics score for 4<sup>th</sup> grade students (see Table 6.3). This scale shows good reliability (Cronbach's  $\alpha = 0.83$ , Martin, Mullis, & Hooper, 2016).

#### 6.2.3.3.3 TIMSS 2015 – 8<sup>th</sup> grade

As above, but using data for each of 40 countries, for 8<sup>th</sup> grade students (see Table 6.3). This scale also shows good reliability (Cronbach's  $\alpha = 0.88$ , Martin et al., 2016).

Table 6.3 Sample sizes and effect sizes (*d*) for sex differences on two measures of Maths ability

Country/ population	PISA (2015)		TIMSS (2015)		TIMSS (2015)	
			4 <sup>th</sup> grade		8 <sup>th</sup> grade	
	Total N	<i>d</i>	Total N	<i>d</i>	Total N	<i>d</i>
Albania	5215	0.05				
Algeria	5519	0.04				
Argentina	1657	-0.09				
Armenia			5383	0.04	5059	0.06
Australia	14530	-0.03	6057	-0.13	10331	-0.04
Austria	7007	-0.12				
Bahrain			4146	0.24	4918	0.24
Belgium	9651	-0.07	5404	-0.11		
Botswana					5964	0.21
Brazil	23141	-0.05				
Bulgaria	5928	0.01	4228	0.07		
Canada	20058	-0.03	12261	-0.15	8754	-0.07
Chile	7053	-0.10	4756	-0.04	4849	-0.19
China	9841	-0.02				
Chinese Taipei	7708	-0.02	4291	-0.11	5711	-0.02
Colombia	11795	-0.05				
Costa Rica	6866	-0.10				
Croatia	5809	-0.07	3985	-0.19		
Cyprus			4125	-0.08		
Czech Republic	6894	-0.04	5202	-0.14		
Denmark	7161	-0.06	3710	-0.08		
Dominican Republic	4740	0.03				
Egypt					7822	0.11
England			4006	-0.07	4814	0.00
Estonia	5587	-0.04				
Finland	5882	0.05	5015	0.11		
Former Yugoslav Republic of Macedonia	5324	0.06				
France	6108	-0.04	4872	-0.09		
Georgia	5316	0.07	3919	0.04		

Germany	6504	-0.09	3942	-0.07		
Greece	5532	0.00				
Hong Kong	5359	-0.01	3600	-0.16		
Hungary	5658	-0.05	5036	-0.09	4893	-0.15
Iceland	3371	0.01				
Indonesia	6513	0.01	4025	0.05		
Iran, Islamic Republic of			3823	0.05	6130	0.01
Ireland	5741	-0.11	4310	-0.08	4704	-0.09
Israel	6598	-0.03			5512	-0.01
Italy	11583	-0.08	4371	-0.31	4481	-0.11
Japan	6647	-0.07	4383	-0.01	4745	0.02
Jordan	7267	0.06			7865	0.21
Kazakhstan			4702	0.06	4887	0.07
Korea, Republic of	5581	0.03	4669	-0.12	5309	-0.04
Kosovo	4826	-0.09				
Kuwait			3587	0.15	4491	0.16
Latvia	4869	0.02				
Lebanon	4546	-0.11			3873	-0.06
Lithuania	6525	0.01	4529	-0.01	4347	-0.02
Luxembourg	5299	-0.11				
Macao	4476	0.11				
Malaysia					9726	0.05
Malta	3634	0.04			3817	0.03
Mexico	7568	-0.05				
Moldova	5325	0.01				
Montenegro	5665	0.00				
Morocco			5068	0.03	13027	0.02
Netherlands	5385	-0.02	4513	-0.16		
New Zealand	4520	-0.06	6322	-0.04	8142	-0.05
Northern Ireland			3115	-0.04		
Norway	5456	0.02	4329	0.04	4697	0.00
Oman			9105	0.20	8883	0.31
Peru	6971	-0.05				
Poland	4478	-0.09	4747	-0.05		
Portugal	7325	-0.06	4693	-0.18		

Qatar	12083	0.08	5194	0.08	5403	0.10
Romania	4876	0.00				
Russian Federation	6036	-0.03	4921	0.01	4780	-0.08
Saudi Arabia			4337	0.51	3759	0.19
Serbia			4036	-0.03		
Singapore	6115	0.00	6517	0.06	6116	0.14
Slovak Republic	6350	-0.03	5773	-0.12		
Slovenia	6406	-0.03	4445	-0.11	4257	-0.05
South Africa					12513	0.00
Spain	6736	-0.11	7764	-0.19		
Sweden	5458	0.01	4132	0.42	4079	-0.06
Switzerland	5860	-0.07				
Thailand	8249	0.01			6482	0.13
Trinidad and Tobago	4692	0.18				
Tunisia	5375	-0.04				
Turkey	5895	-0.02	6456	-0.02	6079	0.07
United Arab Emirates	14167	0.02	21174	0.07	18012	0.08
United Kingdom	14157	-0.05				
United States	5712	-0.05	10029	-0.10	10217	-0.03
Uruguay	6062	-0.08				
Vietnam	5826	0.01				

*Note:* Countries/populations are as reported in PISA (2015) and TIMSS (2015). Negative Cohen's  $d$  means that men have higher scores on the given sex difference.

#### 6.2.4 Comparison data

Locating comparison data provided a particular challenge. There are few sex differences for which there appears to be consensus over their primarily biological or cultural origins, and our choice was driven mainly by feasibility and accessibility of the data in question.

##### 6.2.4.1 *Sex differences in body height*

In order to provide a comparison standard for the psychological sex differences, data regarding a biological sex difference, body height, was also obtained for this study. We attempted to find a single source of data for height differences that included sufficient information to calculate this as Cohen's  $d$ , in order to provide a meaningful comparison for heterogeneity with our

other samples. However, we were unable to locate such a source. We therefore chose a large meta-analysis of height data that included a large number of countries (Gustafsson & Lindenfors, 2009), and used this as a reference point to locate the original data sources listed within this meta-analysis in order to obtain the sample sizes, means, and standard deviations needed (see Table 6.4, and Appendix D for full reference list for this sample). Where more than one source was listed for any given population, we included only the largest sample size. Where heights were reported in inches or millimetres, measurements were converted to centimetres.

Table 6.4 Sample sizes and effect sizes (*d*) for sex differences in height

Country/population	Total N	<i>d</i>	Reference
Anga - Kukukuku	158	-1.88	(Malcolm, 1969)
Arawakan - Central Arawaks	49	-2.29	(Gillin, 1936)
Australian - Aborigine	42	-2.45	(Abbie, 1967)
Australian - South Australian Aborigine	48	-2.80	(Pretty, Henneberg, Lambert, & Prokopec, 1998)
Australian - Yuendumu	65	-1.89	(Barrett & Brown, 1971)
Awin - Awin	133	-1.92	(Hyndman, Ulijaszek, & Lourie, 1989)
Aymara - Aymara (Bolivia)	64	-2.54	(Mueller et al., 1980)
Aymara - Aymara (Chile, altiplana)	160	-2.68	(Mueller et al., 1980)
Aymara - Aymara (Chile, coast)	132	-2.22	(Mueller et al., 1980)
Aymara - Aymara (Chile, sierra)	114	-1.97	(Mueller et al., 1980)
Bane - Bamileke	588	-1.54	(Hiernaux, 1968)
Bane - Bamum	129	-1.15	(Hiernaux, 1968)
Bantu, NW - Duala	125	-2.16	(Hiernaux, 1968)
Bantu, NW - Teke	400	-1.38	(Hiernaux, 1968)
Bantu, SE - Durban Zulus	325	-1.71	(Slome, Gampel, Abramson, & Scotch, 1960)
Basque - Basques espagnols	655	-2.19	(Hiernaux, 1968)
Basque - Basques francais	275	-2.08	(Hiernaux, 1968)
Belgian - Belgium, Brussels	567	-2.09	(Hiernaux, 1968)
Biaka - Binga (Cameroun)	966	-1.30	(Hiernaux, 1968)
Biaka - Binga (Gabon)	96	-1.58	(Hiernaux, 1968)
Biaka - Western pygmies (CAR)	91	-1.05	(Cavalli-Sforza, 1986)



Bougainville SE, Nasioi	108	-2.46	(Friedlaender, 1987)
Caingang - Caingang (Palmas)	62	-1.93	(Neves, Salzano, & Da Rocha, 1985)
Caingang - Caingang (Rio G. do Sul, Parana)	608	-1.76	(Neves et al., 1985)
Chukchi - Chukchi	152	-2.56	(Smirnova, 1979)
Czechoslovakian - Czech lands	7606	-2.13	(Fetter & Hainis, 1962; Hiernaux, 1968)
Czechoslovakian - Slovakia	344	-1.88	(Fetter & Hainis, 1962)
Druse - Druse	295	-3.55	(Shanklin & Izzeddin, 1937)
Dutch - Hollandais du Nord	130	-2.54	(Hiernaux, 1968)
English - Great Britain	9863	-1.99	(Rosenbaum, Skinner, Knight, & Garrow, 1985)
Eskimo (Alaskan) - Alaskan	225	-2.02	(Auger et al., 1980)
Eskimo (Canadian) - Igloodik (Foxe Basin)	248	-2.07	(Auger et al., 1980)
Eskimo (Canadian) - Igloodik Eskimo	24	-1.51	(Shephard, 1974)
Eskimo (Canadian) - Labrador Inuit	136	-2.07	(Skeller, 1954)
Eskimo (Greenland) - Angmagsalik Inuit	369	-1.59	(Skeller, 1954)
Eskimo (Greenland) - West Greenland	86	-2.35	(Auger et al., 1980)
Estonians	5065	-3.11	(Hiernaux, 1968)
Fiji - Fiji	346	-2.03	(Hawley & Jansen, 1971)
Fiji - Fiji-Melanesian	25	-6.08	(Clegg, 1989)
Finnish - Finlandais de Botnic	946	-1.92	(Hiernaux, 1968)
French - Francais	120	-1.37	(Hiernaux, 1968)
Fulani - Peul du Niger	83	-1.80	(Hiernaux, 1968)
Fulani - Peul du Sud-Cameroun	99	-2.27	(Hiernaux, 1968)
German - Allemands du Centre	858	-1.96	(Hiernaux, 1968)
Guinea - Bedik - Bassari	219	-1.56	(Hiernaux, 1968)
Gur - Mossi (Donse)	107	-1.60	(Froment & Hiernaux, 1984)
Gur - Mossi (Kokologo)	146	-1.72	(Froment & Hiernaux, 1984)
Hausa - Hausa du Cameroun	85	-1.51	(Hiernaux, 1968)
Hausa - Hausa du Niger	317	-1.63	(Hiernaux, 1968)

Hazda	148	-1.76	(Barnicot, Bennett, Woodburn, Pilkington, & Antonis, 1972)
Hungary - Hongrois	171	-1.91	(Hiernaux, 1968)
Ibo - Ibo orientaux	122	-1.70	(Hiernaux, 1968)
Iranian (E + W) - South Iranian	1846	-1.23	(Ayatollahi & Carpenter, 1993)
Irish - Irish	10703	-1.98	(Hooton & Dupertuis, 1955)
Koryak - Koryak	306	-2.33	(Brodsky, 1906)
Kurdish - Kurds (Iraq)	628	-2.28	(Field, 1952)
Lapp Finnish - Finnish Lapps	488	-1.91	(Auger et al., 1980)
Liberia - Kru - Kran	200	-1.72	(Hiernaux, 1968)
Luangiua - Ontong Java	168	-1.78	(Friedlaender, 1987)
Macushi - Macushi	40	-2.72	(Farabee, 1924)
Makiritare - Maquiritare	50	-1.99	(Díaz Ungría, 1969)
Malaita - Kwaio	213	-2.39	(Friedlaender, 1987)
Malaita - Lau (Malaita)	137	-2.13	(Friedlaender, 1987)
Mapuche Araucanian - Araucanian	146	-1.91	(Valenzuela, Rothhammer, & Chakraborty, 1978)
Mayan E - Quiche	200	-2.22	(Comas, 1971)
Mbuti - Epulu	101	-1.23	(Cavalli-Sforza, 1986)
Mbuti - Ituri (Eastern Pygmies)	109	-2.10	(Cavalli-Sforza, 1986)
Mbuti - Mbuti (Congo Leopoldville)	892	-1.14	(Hiernaux, 1968)
Na-Dene (Canadian) - Chilcotin Athapascan	91	-2.86	(Birkbeck, Lee, Myers, & Alfred, 1971)
Na-Dene (Canadian) - Chippewyan	64	-2.52	(Skeller, 1954)
Na-Dene (Canadian) - Dogrib	157	-1.95	(Szathmary & Holt, 1983)
Nentsy - Wood Nenetz	92	-1.95	(Smirnova, 1979)
New Guinea - Highland E - Auyana	290	-1.23	(Littlewood, 1972)
New Guinea - Highland E - Gadsup	267	-1.59	(Littlewood, 1972)
New Guinea - Highland E - Tairora	343	-1.03	(Littlewood, 1972)
Nilotic - Maasai	268	-1.93	(Sellen, 1995)
Nilotic - Turkana	233	-1.47	(Little & Johnson Jr, 1986; Sellen, 1995)
Norwegians - Norwegians	11967	-1.84	(Hiernaux, 1968)

Ok - Mountain Ok (Papua New Guinea)	291	-1.03	(Schwartz, Brumbaugh, & Chiu, 1987)
Peul - Fulakunda (Peul) du Badyar	200	-2.10	(Hiernaux, 1968)
Portuguese - Portugais	350	-2.07	(Hiernaux, 1968)
Pygmoid - Twa (Rwanda)	185	-1.44	(Hiernaux, 1968)
Quechua - Quechua	282	-2.51	(Rouma, 1933)
Quechua - Quechua (highland)	120	-2.41	(Frisancho, Borkan, & Klayman, 1975)
Quechua - Quechua (lowland)	117	-2.53	(Frisancho et al., 1975)
Quechua - Quechua (Nunua, Peru)	100	-4.16	(Frisancho & Baker, 1970)
Russians	542	-1.92	(Hiernaux, 1968)
Samoa - Salammumu (Western Samoa)	245	-2.11	(Bindon & Baker, 1985)
San - Kung (Bochimans, Af. S. O.)	135	-1.82	(Hiernaux, 1968)
Sara - Sara Madjingay	751	-1.47	(Crogner, 1979)
Saudi - Saudi (Highlanders)	437	-2.07	(Khalid, 1995)
Saudi - Saudi (Lowlanders)	468	-1.93	(Khalid, 1995)
Sea Dayak - Iban	84	-2.45	(Strickland & Ulijaszek, 1993)
Society - Society Islands	153	-1.82	(Shapiro, 1930)
Solomon Islands - Bougainville W - Nagovisi	210	-1.96	(Harrison, 1977)
Solomon Islands - Malaita - Baegu	237	-2.47	(Harrison, 1977)
South Africa - Bantu, SE - Venda	224	-2.31	(Hiernaux, 1968)
South Chinese - Hong Kong	789	-2.20	(Chang, 1969)
Swedish - Suedois de Runo	152	-2.40	(Hiernaux, 1968)
Taiwan - Ami - Ami	240	-1.52	(Chai, 1967)
Taiwan - Atayal - Atayal	243	-2.02	(Chai, 1967)
Taiwan - Bunun - Bunun	206	-2.07	(Chai, 1967)
Taiwan - Paiwan - Paiwan	277	-1.70	(Chai, 1967)
Tolai - Tolai	103	-0.90	(Champness, Bradley, & Walsh, 1963)
Trio - Trio (Surinam)	257	-2.05	(Glanville & Geerdink, 1970)
Tungus - Evenki Reindeer Herders	39	-2.38	(Leonard, Katzmarzyk, Comuzzie, Crawford, & Sukernik, 1994)

Uganda - Bantu, NE - Ganda	503	-1.85	(Hiernaux, 1968)
Ukrainians	450	-1.65	(Hiernaux, 1968)
Venezuela - Mari - Motilon	74	-1.05	(Fleury-Cuello, 1953)
Volta - Agni	120	-1.84	(Hiernaux, 1968)
Wajana - Wajana (Surinam)	166	-2.28	(Glanville & Geerdink, 1970)
Warau - Warao	340	-2.28	(Fleischman, 1980)
Xavante - Xavante	81	-2.89	(Niswander, Keiter, & Neel, 1967)
Yanomama - Yanomama	139	-2.35	(Neves et al., 1985)
Yanomama - Yanomamo	576	-2.33	(Spielman et al., 1972)
Yoruba - Akufo (Yoruba)	545	-5.00	(Janes, 1970)
Yugoslavian - Yugoslaves	192	-2.33	(Hiernaux, 1968)

*Note:* Countries/populations are as reported in Gustafsson and Lindenfors (2009). Negative Cohen's *d* means that men are taller than women.

#### 6.2.4.2 *Sex difference in 2d:4d digit ratio*

The majority of data for this effect were based on a meta-analysis by Hönekopp and Watson (2010), which listed sex differences separately for studies conducted in multiple different countries. For each country, we retained the largest study reported in the meta-analysis that examined an adult sample, and used a reliable method of measurement for 2d:4d (Hönekopp & Watson, 2010).

Following extraction of data from this meta-analysis, due to the relatively small number of countries included (compared to the other sex differences in our sample) we attempted to also obtain more recent papers which reported 2d:4d ratio for samples from countries not included in the meta-analysis. We therefore performed a search on Web of Science in November 2018, using the same search terms as the meta-analysis ("2d:4d" OR "digit ratio" OR "finger length") from 2009 (when the meta-analysis was submitted for publication) to present. This returned 809 search results.

We used the following inclusion criteria to filter the search results: i) studies must report sex-differences in 2d:4d ratio, or sufficient data for this to be calculated. ii) sex differences must be reported on adults, and males and females in the study must be of a similar age. iii) data must be available for a specific country not included in the original meta-analysis. iv) the sample must be ethnically homogenous. v) for practical reasons, the study must be available in

English. In addition to this, we excluded studies where finger lengths were measured by participants themselves, or measured from drawn hand outlines, as these two measurement methods have been found to be less accurate than experts' measures performed directly on the hands or their photocopies (Hönekopp & Watson, 2010). Finally, paralleling the other included datasets, only the largest study for adult samples for any given country was included in our sample. These were then added to the studies included from the meta-analysis (see Table 6.5).

Our final sample includes data from 18 countries from the meta-analysis (Hönekopp & Watson, 2010) and an additional 24 countries from later studies (Agnihotri, Jowaheer, & Soodeen-Lalloo, 2015; Almasry, El Domiaty, Algaidi, Elbastawisy, & Safwat, 2011; Andrievskaya & Semenova, 2017; Apicella, Tobolsky, Marlowe, & Miller, 2016; Aycinena, Baltaduonis, & Rentschler, 2014; Barel & Tzischinsky, 2017; Camacho-Hernandez et al., 2018; Chicaiza-Becerra & Garcia-Molina, 2017; Darnai et al., 2016; Galeta, Bruzek, & Laznickova-Galetova, 2014; Gonçalves, Coelho, Machado, & Rocha, 2017; Hsu et al., 2015; Jeon et al., 2016; Joyce et al., 2014; Kyriakidis, Papaioannidou, Pantelidou, Kalles, & Gemitzis, 2010; Lu et al., 2017; Marczak, Misiak, Sorokowska, & Sorokowski, 2018; Nye, Androuschak, Desierto, Jones, & Yudkevich, 2012; Oyeyemi et al., 2014; Rivas et al., 2014; Sorokowski, Sorokowska, Danel, Mberira, & Pokrywka, 2012; Wakabayashi & Nakazawa, 2010; Weisman et al., 2015; Zhao, Li, Yu, & Zheng, 2012). Only right hand 2d:4d was used.

Table 6.5 Sample sizes and effect sizes (*d*) for 2d:4d digit ratio

Country/population	Total N	<i>d</i>	Reference
Australia*	1093	0.53	
Austria*	1118	0.49	
Belgium*	580	0.50	
Brazil	200	0.80	(Rivas et al., 2014)
Canada*	636	0.51	
China - Han Chinese	1376	0.61	(Weisman et al., 2015)
China - Hani ethnicity	140	0.50	(Zhao, Yu, Zhang, & Zheng, 2013)
China - Hui ethnicity	347	0.55	(Lu et al., 2017)
Columbia	123	0.73	(Chicaiza-Becerra & Garcia-Molina, 2017)
Czech Republic*	300	0.42	
France	100	0.83	(Galeta et al., 2014)
Germany*	735	0.19	

Greece	80	0.92	(Kyriakidis et al., 2010)
Guatemala	219	0.67	(Aycinena et al., 2014)
Hungary	75	0.31	(Darnai et al., 2016)
India*	160	0.25	
Indonesia - Yali population	79	-0.23	(Marczak et al., 2018)
Ireland	148	0.37	(Joyce et al., 2014)
Israel	80	0.82	(Barel & Tzischinsky, 2017)
Italy*	292	0.45	
Jamaica*	146	0.33	
Japan	348	0.67	(Wakabayashi & Nakazawa, 2010)
Korea	691	0.83	(Jeon et al., 2016)
Lithuania*	109	0.76	
Mauritius	200	0.63	(Agnihotri et al., 2015)
Namibia	97	0.84	(Sorokowski et al., 2012)
Nigeria	801	0.89	(Oyeyemi et al., 2014)
Philippines	123	0.70	(Nye et al., 2012)
Poland*	213	0.19	
Portugal	175	0.26	(Gonçalves et al., 2017)
Russia	4407	0.06	(Andrievskaya & Semenova, 2017)
Saudi Arabia	560	-0.25	(Almasry et al., 2011)
South Africa*	138	0.33	
Spain*	88	0.06	
Sweden*	96	0.88	
Switzerland*	389	0.57	
Taiwan	316	0.29	(Hsu et al., 2015)
Tanzania - Hadza population	152	-0.55	(Apicella et al., 2016)
Tenerife	164	0.31	(Camacho-Hernandez et al., 2018)
Turkey*	183	0.82	
United Kingdom*	680	0.27	
United States*	1140	0.41	

*Note:* Countries/populations are as reported in original papers, or in meta-analysis (Hönekopp & Watson, 2010). Negative Cohen's *d* means that men have higher 2d:4d ratio than women.

\* Reported in Hönekopp & Watson (2010)

#### 6.2.4.3 *Cultural sex differences*

In order to provide a comparison standard for the psychological sex differences, we collated data for sex differences that are thought to be primarily cultural in nature. Firstly, we used data from the Programme for International Student Assessment (PISA), 2006 (*PISA 2015 Database*, 2015). Following a similar analysis undertaken in a cross-cultural meta-analysis by Mann, Legewie, & DiPrete (2015), we utilised questions around students' interest in science, and careers in STEM (science, technology, engineering and mathematics), which were last asked in the 2006 version of this survey.

Specifically, for interest in science, we used the scores from question 21 of the student questionnaire. This question asked students to rate their interest in eight science topics, including topics in physics, chemistry, plant biology, human biology, astronomy, and geology, as well as how scientists design experiments, and what is required for scientific explanations. Students rated their interest from 1 (high interest) to 4 (no interest). We summed responses to create a total score for interest in science, which could range from 8 (highest interest in science) to 32 (lowest interest in science). We then obtained the mean and SD for males and females, for each country, in order to calculate Cohen's  $d$  for the sex difference. We included data from the 57 countries available from this sample (see

Table 6.6).

For future career (interest in a career in STEM), we used scores from question 29 of the student questionnaire. This question asked students to rate their level of agreement with 4 statements relating to STEM aspirations: would like to work in a career involving STEM; would like to study STEM after compulsory education (secondary school); would like to spend life doing advanced STEM; would like to work on STEM projects as an adult. We summed responses to create a total score for interest in STEM, which could range from 4 (highest interest in working in STEM) to 16 (lowest interest in working in STEM). Male and female mean scores (and standard deviations) were then calculated for each country. We included data from the 57 countries available from this sample (see



Table 6.6).

Table 6.6 Sample sizes and effect sizes ( $d$ ) for sex differences in interest in science, and future career (interest in a career in STEM)

Country/ population	Interest in science		Future career	
	Total N	$d$	Total N	$d$
Argentina	4339	0.11	4063	0.00
Australia	14170	-0.03	13839	0.08
Austria	4927	-0.06	4814	0.10
Azerbaijan	5184	0.11	4637	0.01
Belgium	8857	-0.07	8503	0.19
Brazil	9295	0.09	8517	0.05
Bulgaria	4498	0.11	4262	-0.11
Canada	22646	-0.02	21879	-0.02
Chile	5233	0.10	5046	0.01
China	9405	-0.13	9326	0.32
Chinese Taipei	8812	-0.20	8775	0.63
Colombia	4478	0.22	4370	-0.04
Croatia	5213	0.05	5127	-0.03
Czech Republic	5932	0.18	5719	-0.21
Denmark	4532	0.18	4271	-0.07
Estonia	4865	0.05	4811	-0.04
Finland	4714	0.07	4640	-0.11
France	4716	-0.05	4535	0.18
Germany	4891	-0.02	4495	0.22
Greece	4873	0.08	4778	0.34
Hungary	4490	0.10	4431	-0.02
Iceland	3789	0.03	3700	0.29
Indonesia	10647	0.18	10413	-0.09
Ireland	4585	-0.07	4425	-0.09
Israel	4584	-0.06	4067	0.15
Italy	21773	-0.14	21386	0.20
Japan	5952	-0.13	5926	0.45
Jordan	6509	0.18	6330	0.20
Korea	5176	-0.12	5164	0.29
Kyrgyzstan	5904	0.24	5363	-0.05
Latvia	4719	0.03	4639	-0.02

Liechtenstein	339	0.06	331	0.19
Lithuania	4744	0.19	4693	-0.09
Luxembourg	4567	-0.01	4458	0.03
Mexico	30971	0.04	30550	0.18
Montenegro	4455	0.24	4267	-0.03
Netherlands	4871	-0.14	4616	0.28
New Zealand	4823	-0.04	4687	0.04
Norway	4692	0.07	4461	0.16
Poland	5547	-0.03	5497	-0.14
Portugal	5109	0.01	5037	0.04
Qatar	6265	-0.06	5766	0.33
Romania	5118	0.09	5025	0.03
Russian Federation	5799	0.09	5718	0.13
Serbia	4798	-0.02	4698	0.14
Slovak Republic	4731	-0.13	4659	-0.12
Slovenia	6595	-0.13	6219	0.00
Spain	19604	-0.01	19312	0.08
Sweden	4443	0.15	4307	0.08
Switzerland	12192	-0.08	11996	0.12
Thailand	6192	0.26	6135	-0.03
Tunisia	4640	0.07	4464	0.07
Turkey	4942	-0.11	4846	0.11
United Kingdom	13152	-0.07	12743	0.15
United States	5610	-0.20	5486	0.14
Uruguay	4839	0.03	4502	-0.11

*Note:* Countries/populations are as reported in PISA (2006). Negative Cohen's *d* means that boys have higher scores on the given sex difference

In addition to this, we also included a meta-analysis which examined adults' attitudes towards gender roles and behaviours (Stickney & Konrad, 2007). This utilised data from the International Social Survey Program (ISSP), 2002, a cross-national collaboration of institutions (e.g. universities, survey agencies) that conducts surveys on social science related topics. Attitudes toward gender roles were assessed using four items from this survey, which were determined to be most comparable across cultures. The four items were: a pre-school child is likely to suffer if his or her mother works; all in all, family life suffers when the woman has a full-time job; a job is all right, but what most women really want is a home and children; a

husband's job is to earn money, a wife's job is to look after the home and family (Stickney & Konrad, 2007). These four items accounted for 55.6% variance in items (based on principal components analysis), and had a reliability of .73. Items were assessed on a 5-point Likert scale from 1 (strongly agree) to 5 (strongly disagree), with higher scores relating to more egalitarian attitudes. Gender-Role Attitude Index scores were created by averaging scores from the four items. We included data from the 28 countries available from this sample (see Table 6.7).

Table 6.7 Sample sizes and effect sizes (*d*) for sex differences in attitudes toward gender roles and behaviours

<b>Country/population</b>	<b>Total N</b>	<b><i>d</i></b>
Australia	204	0.39
Austria	372	0.56
Brazil	282	0.22
Chile	328	0.02
Cyprus	378	0.62
Czech Republic	147	0.2
Denmark	387	0.29
Finland	373	0.42
France	540	0.36
Germany	211	0.55
Great Britain	614	0.34
Hungary	185	0.06
Israel	336	0.33
Japan	314	0.27
Latvia	257	0.08
Mexico	256	-0.04
Northern Ireland	185	0.05
Philippines	439	0
Poland	320	0.35
Portugal	266	0.25
Russia	378	0.15
Slovak Republic	388	0.19
Slovenia	208	0.05
Spain	414	0.47
Sweden	303	0.28

Switzerland	193	0.39
Taiwan	559	0.5
United States of America	316	0.45

*Note:* Countries/populations are as reported in Stickney and Konrad (2007). Negative Cohen's *d* means that men have higher scores than women

#### 6.2.5 Data extraction and analysis

Overall, we had cross-cultural data on 15 sex differences. After extracting the relevant data for each of the samples for each sex difference, we calculated Cohen's *d* for each individual sample using the *escalc* function in Metafor in R. We then ran a meta-analysis to calculate the mean effect size and a measure of heterogeneity (*T*) using the DL estimator in Metafor (additional results for the HS and PM estimators are reported in Appendix D). Using the raw data in this way allows for the variance to be calculated more accurately when group sizes differ (see section 2.1.6). In order to establish whether this created any significant differences in mean effect size or heterogeneity, we also conducted the main meta-analyses using Cohen's *d* and total group size (the approach that we had to use for Study 1a and Study 2). For details, see Appendix D.

We believe that obtaining heterogeneity estimates for our sample of 10 psychological sex differences is sufficient to make an assessment of the usefulness of (cross-cultural) heterogeneity in the context of sex differences. We will also have heterogeneity estimates for 2 biological sex differences and 3 cultural sex differences by way of a comparison sample. However, making a simple direct comparison of heterogeneity levels between our psychological sex differences sample and other samples is problematic, because of the relationship between *d* and *T* (see Study 1a). We therefore use our previous analyses from Study 1a to provide a useful alternative to direct comparison. In Study 1a, we were able to predict levels of heterogeneity based on the effect size from any given meta-analysis, or based on the results from multiple close replications. The sex differences included in the present study were all based on studies which used either the same, or very similar, measures of both the independent and dependent variables. We should therefore expect that very little heterogeneity should be introduced from variations in methodology, and that the studies from multiple countries could therefore be considered a form of close replication. We used the regression equation for close replications from this Study 1a, based on 57 close replication studies, to predict the level of heterogeneity that we would expect for each sex difference ( $T = 0.067 + 0.146d$ ). We then calculated the residuals (actual heterogeneity minus predicted

heterogeneity) to give an estimate of the magnitude of any ‘excess’ heterogeneity over and above this expected amount, or an estimate of how much lower heterogeneity might be than expected. Following study 1a, we also calculated the Flop Index (*FI*) for each of the effects. *FI* tells us the proportion of results from replications that would be expected to find effects in the opposite direction to those found by a meta-analysis. Large, positive residuals indicate large excess heterogeneity for the given effect size. If accompanied by high *FI*, this might be seen as plausible evidence for the social structuralist perspective (see Figure 6.1, right panel). Large, negative residuals indicate lower than expected heterogeneity for the given effect size. If this is accompanied by low *FI*, this might be seen as plausible evidence for the evolutionary perspective, as it would suggest that a sex difference would be plausible in the EEA (see Figure 6.1, left panel).

### 6.3 Results

#### 6.3.1 Comparison data

Table 6.8 shows descriptive statistics for effect sizes and heterogeneity of each of the sex differences, including the data for the comparison effects. As can be seen from the table, our expectations regarding the levels of heterogeneity for our comparison effects were not borne out by the data.

The data for sex differences thought to be primarily underpinned by cultural influences (Interest in science, Career in STEM, and Gender Role Attitudes) showed low absolute levels of heterogeneity, all of which did not differ greatly from the levels expected under conditions of close replication (see residuals column in Table 6.8 for differences between actual and expected levels of heterogeneity). However, it is worth noting the *Flop Index (FI)* for each of these effects. Remember that *FI* tells us the proportion of studies expected to have true effects in the opposite direction to those found by the meta-analysis. *FI* is affected by both the effect size and its level of heterogeneity. For both Interest in Science and Career in STEM, very small effects were found showing that men score slightly more highly on these measures than women. This is also reflected in high *FI* for both effects (43% for Interest in Science, and 29% for Career in STEM), suggesting that a high proportion of studies would be expected to have true effects in the opposite direction. There is therefore not a strong argument for a consistent underlying sex difference in either measure. For Gender role attitudes, there appears to be a medium sex difference ( $d = 0.298$ ) showing that women have more egalitarian gender role attitudes than men. The level of heterogeneity is again relatively small, suggesting that this is a

relatively stable effect with a low proportion of studies likely to run counter to expectation ( $FI = 2\%$ ).

In contrast to this, height shows a particularly strong effect (men are, on average, taller than women), that is consistent with no studies expected to run against this finding ( $FI = 0.0\%$ ), however heterogeneity is larger in this sample than we might expect. Why might this be the case? For height data, a particularly large range of dates of measurement were available for the different populations, ranging from samples measured in the late 19<sup>th</sup> century to those measured in the 1990s (see Appendix D). In addition to this, while all data were collected from adult samples, samples of different ages were measured. While the direction of this sex difference is clearly consistent, and this difference is widely believed to be underpinned by biological influences (Archer, 2019; Gray & Wolfe, 1980), it is possible that additional sources of heterogeneity could have been introduced. For example, Gleeson and Kushnick (2018) suggest that food security and societal status can affect stature, and found that higher female status is associated with less sexual dimorphism, particularly when food resources are secure. They reason that when resources are low, female mate preference for masculinised males (with height as one component of this) is higher. These factors are impossible to control for, particularly given the diverse date range of data collection and variation in ages of participants when measured.

For 2d:4d digit ratio, the picture is slightly less clear. A medium effect was found ( $d = 0.46$ ) showing that women tend to have higher 2d:4d ratio than men. However, levels of heterogeneity were more than twice that expected from close replications. In addition to this,  $FI$  shows that 5% of replications would be expected to run in the opposite direction. Similarly to the height data, this sex difference could be suffering from the same difficulty in terms of measurement. Data for this effect were collected from a meta-analysis and a number of additional studies from 24 other countries.

Table 6.8 Descriptive statistics for effect size ( $d$ ) and heterogeneity ( $T$ ) of each sex difference

Sex difference (no. countries)	$d$	$T$	Predicted $T$	Residual $T$	Flop Index ( $FI$ )
Mate preference: financial (37)	0.738	0.328	0.175	0.153	0.01
Mate preference: ambition (37)	0.415	0.269	0.128	0.142	0.06
Mate preference: age difference (37)	2.181	0.548	0.385	0.163	0.00
Mate preference: looks (37)	- 0.590	0.195	0.153	0.042	0.00
Mate preference: chastity (37)	- 0.323	0.220	0.114	0.106	0.07
Mental rotation: line angle (53)	- 0.510	0.077	0.141	- 0.064	0.00
Mental rotation task (52)	- 0.462	0.042	0.134	- 0.093	0.00
Mathematics ability at 15: PISA 2015 (69)	- 0.023	0.048	0.070	- 0.022	0.32
Mathematics ability 4 <sup>th</sup> grade: TIMSS 2015 (48)	- 0.018	0.138	0.070	0.069	0.45
Mathematics ability 8 <sup>th</sup> grade: TIMSS 2015 (40)	0.027	0.104	0.070	0.033	0.40
Biological sex differences					
Height (117)	- 2.009	0.499	0.360	0.139	0.00
2d:4d right hand (42)	0.462	0.279	0.135	0.144	0.05
Cultural sex differences					
Interest in Science: PISA 2006 (57)	- 0.019	0.112	0.070	0.042	0.43
Career in STEM: PISA 2006 (57)	- 0.098	0.154	0.080	0.075	0.29
Gender role attitudes: ISSP 2002 (28)	0.298	0.140	0.109	0.032	0.02

*Note:* negative Cohen's  $d$  means that men have higher scores on the given sex difference (e.g. men are taller than women, which is reflected in a large, negative Cohen's  $d$ ). For interpretation of residuals and  $FI$ , see section 6.2.5.



### 6.3.2 Psychological sex differences

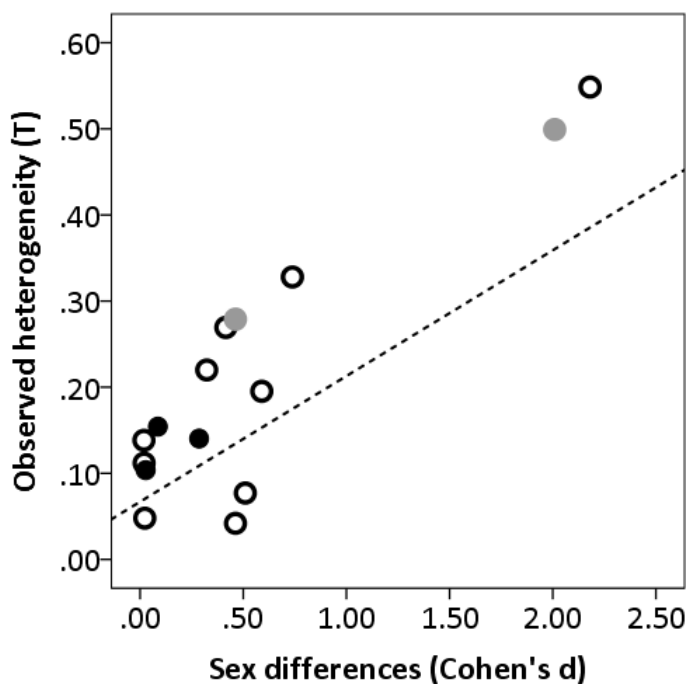
Most of the measures of mate preference showed medium to large effect sizes, with resulting low *FI* (the highest *FI* for this subset was 7%). However, most also showed a substantial amount of excess heterogeneity, in some cases more than double that which would be expected for close replications for these effect sizes. Additional sources of heterogeneity in these samples could be due to the type of measure being used, namely self-report questionnaires, which were also translated into multiple different languages to assess samples in 37 different countries (Buss, 1989). While the authors took care to ensure accuracy of translation as far as possible, self-report measures could be particularly unreliable and questions variable in how people could interpret them (Beaton, Bombardier, Guillemin, & Ferraz, 2000). The study used a newly developed measure that had not been validated previously. Additionally, the author notes that sampling techniques varied between countries, that sample representativeness differs, and that the same protocol was not able to be sent to the researchers in each country (Buss, 1989). All of these potential differences in data collection could have increased observed levels of heterogeneity.

In contrast to this, measures of mental rotation (both the task and the line angle measure) showed medium effect sizes (Cohen's *d* of - 0.46 and - 0.51, respectively). Both effects had heterogeneity at levels slightly lower than that expected from close replications. Both showed consistency in the direction of effects anticipated in future (*FI* = 0%), suggesting that males consistently perform better at mental rotation tasks than females.

Finally, measures of mathematics ability show inconsistency in the direction of the effect, with PISA and TIMSS 4<sup>th</sup> grade both suggesting a very small (*d* = -0.02 in each case) advantage for males, and TIMSS 8<sup>th</sup> grade suggesting a similarly small (*d* = 0.03) advantage for females. Levels of heterogeneity for all three measures were consistently low, and broadly in line with that expected from close replications. Perhaps unsurprisingly given that all have effect sizes close to zero, all three measures have high *FI* (ranging from 0.32 to 0.45), showing the lack of a consistent sex difference in mathematics ability. Both TIMSS and PISA have been conducted many times previously, allowing for adjustment and improvement of the measures over time to ensure good reliability (see sections 6.2.3.3.1, 6.2.3.3.2, and 6.2.3.3.3) and reduce error. In addition to this, TIMSS and PISA are both conducted on particularly large samples for each country. This allows more accurate assessment of sampling error, and thus more accurate assessment of heterogeneity.

### 6.3.3 Exploratory analyses

In both study 1a and study 2, we found that mean effect size ( $d$ ) and heterogeneity ( $T$ ) correlated across multiple meta-analyses, with larger mean effects correlating with higher levels of heterogeneity. Correlating mean  $T$  with  $d$  for all sex differences (including comparison data sets), resulted in  $r = 0.89$ ,  $p < .001$ . This relationship thus appears to be robust across different samples. In addition to this, in study 1a, we calculated a regression equation from close replications to estimate heterogeneity for close replication studies, which resulted in  $T = 0.067 + 0.146d$  (dotted line in Figure 6.3). Similarly to this, for the current sample which included 15 large close replications, the regression equation was  $T = 0.101 + 0.202d$  ( $R = 0.887$ ).



empirical key arguments made in support of each (for evolutionary perspective, the consistency argument; for social structuralist perspective, the covariation argument) are conceptually flawed. We suggested that variability of effect sizes is key to the interpretation of the data presented in support of each perspective. Therefore, we examined heterogeneity in three domains of psychological sex differences (mental rotation, mathematics ability, and mate preferences) to attempt to add a new perspective to this debate, and compared these to two sex differences that are widely agreed to be biological in origin (height, and 2d:4d digit ratio), and three sex differences thought to be primarily cultural in origin (interest in science, careers in STEM, and attitudes towards gender roles and behaviours). Contrary to our expectations, our comparison effects for biological sex differences showed large cross-cultural heterogeneity relative to the size of the sex difference, and two of our comparison effects for cultural sex differences showed surprisingly low levels of heterogeneity. Levels of heterogeneity were as expected for close replications for mental rotation and mathematics ability, but showed considerable variation over and above that expected for close replications for mate preference outcomes.

#### 6.4.1 Heterogeneity could be introduced through use of unreliable forms of measurement

In study 1, we suggested that key concepts of measurement, such as reliability and validity, were often given great consideration in measurement of the dependent variable in psychological studies. In the current study, we are not able to directly measure the effect of unreliable sources of measurement on levels of heterogeneity. However, based on examples including unvalidated measures (relating to measurement of mate preference, Buss, 1989) compared with measures that have been extensively tested to be highly reliable and valid across multiple cultures (mathematics ability assessed using TIMSS or PISA, *PISA 2015 Database*, 2015; *TIMSS 2015 International Database*, 2015), this could suggest that unreliability of measurement relating to lack of psychometric validity of measures employed to assess the dependent variable can inflate observed levels of heterogeneity. It is also possible that the use of self-report measures, for example those employed by Buss (1989) could introduce heterogeneity in a way that more objective measures, such as those of ability where there are correct and incorrect answers, would not.

Issues of measurement could also apply to data used for the comparisons relating to biological sex differences. For 2d:4d digit ratio, although we excluded studies employing less reliable forms of measurement of finger length (for example drawn outlines of hands, self-measurement), we included multiple more reliable forms of measurement conducted by

experts, both direct (using analogue or digital calipers) and indirect (scans, photographs or photocopies of hands). Although differences between these forms of measurement might not be large, this could still introduce a source of heterogeneity due to potential differences in accuracy between these measures.

#### 6.4.2 Inferences regarding the origins of sex differences

With the data we have presented here, unfortunately our approach does not work as well as we had anticipated (compare our results in Figure 6.3 with our expected pattern of results in Figure 6.2). In large part this is because our sample of biological sex differences showed higher heterogeneity than expected. We can, however, make some comparisons within our sample of psychological sex differences. For example, consider our mental rotation sex differences (see Table 6.8). These show negative residuals (less heterogeneity than expected) and *FI* of 0.0%. This suggests that mental rotation is likely to have biological underpinnings, as it is consistent in direction and low in variability, meaning that it is likely to remain in the EEA. Contrast this with mate preferences in ambition (see Table 6.8). This shows positive residual heterogeneity, and a *FI* of 6.0% suggesting that the effect could reverse. The evidence for an evolved component of sex differences appears to be stronger for mental rotation than it does for mate preferences in ambition.

#### 6.4.3 Comparisons with the samples from study 1a and study 2

We found a significant relationship between *T* and *d* across samples from the current study, study 1a, and study 2. In addition to this, we found similarities between the sample of studies in the current study, which could all be considered to be close replications of the same effect, and the sample of close replications examined in study 1a. We calculated similar regression equations for the prediction of *T* from *d* in each sample. This relationship appears to be robust across a variety of different samples, whether based on close or conceptual replications.

#### 6.4.4 Strengths and limitations

The current study employed multiple large data sets across multiple different cultures (up to 69 countries) involving large numbers of participants to attempt to make accurate assessments of observed heterogeneity in sex differences in several different domains. We identified key flaws in the central arguments about the origin of sex differences, and identified that

heterogeneity could make an important contribution to add to assessment of the evidence for each perspective.

Although we believe that the heterogeneity perspective has a lot of potential, a key limitation to this study was that the data for comparison from biological sex differences showed large residual heterogeneity. This limitation in comparison data could also have led to suboptimal use of heterogeneity (through our use of residual heterogeneity). More comparison data could be helpful in this regard.

#### 6.4.5 Conclusions and future directions

Heterogeneity appears to offer a promising addition to existing perspectives on the origins of sex differences. In future it would be interesting to examine a number of other psychological sex differences, and obtain further data on both biological sex differences and culturally based sex differences to provide a more comprehensive comparison sample. For example, data on differences in employment levels or income were not possible to obtain within the time constraints of the current study, but are collated across the world and could provide an interesting form of comparison.

## 7 General Discussion

The replication crisis in psychology has led to extensive debate regarding the reasons for difficulties with replication of research findings (Zwaan, Etz, Lucas, & Donnellan, 2018). This may be one of the reasons that heterogeneity of empirical results has received increasing attention in the last few years (Kenny & Judd, 2019; McShane & Böckenholt, 2014; Shrout & Rodgers, 2018; Stanley, Carter, & Doucouliagos, 2018; Van Erp, Verhagen, Grasman, & Wagenmakers, 2017). Here we provide a brief summary and discussion of the key findings of each of the studies in this thesis and the contribution this project makes to the replication debate.

### 7.1 Study 1a: Descriptive overview of heterogeneity

In study 1a we extended the existing reviews of heterogeneity in more specific examples by comparing heterogeneity between close and conceptual replications, published in any journal and across three psychological subfields (cognitive, organisational, and social psychology). Using Cohen's  $d$  as our measure of effect size, we found high levels of heterogeneity for our sample of 150 meta-analyses, which can be considered to be compilations of related conceptual replications (Schmidt & Oh, 2016). This level of heterogeneity was supported by the findings of study 2, in which the control sample of meta-analyses had a very similar level of heterogeneity, suggesting that this level of heterogeneity could be common across psychology disciplines, at least those which utilise effect sizes of standardised mean difference or correlations. In contrast to this, across 57 multiple close replications examined in study 1a, including many labs and registered replication reports, we found low levels of heterogeneity. We found no difference in heterogeneity between the three sub-fields (cognitive, organisational, and social psychology). Our exploratory analyses looked at some possible reasons for such high levels of heterogeneity in meta-analyses. We examined the role of moderators, and explored other relationships such as the influence of effect size, median sample size, and the maturity of a research field. We found a strong relationship between Cohen's  $d$  and heterogeneity, showing that as effect sizes increase, levels of heterogeneity also increase. This relationship was consistent across all studies in this project, and allows us to predict levels of heterogeneity for any given phenomenon for which a measure of effect size can be estimated. This prediction model facilitates the inclusion of heterogeneity in power calculations, allowing more accurate determination of sample sizes needed for sufficiently powered experiments (McShane & Böckenholt, 2014).

### 7.1.1 Implications of heterogeneity

Where heterogeneity is high, the result of the next study is relatively uncertain. Our heterogeneity results therefore suggest that the effectiveness of a new practical application might be difficult to predict with any accuracy, and that we could expect a proportion of results to go against the expected direction. However, all else being equal, we would expect a practical application to have a greater chance of success when the average effect size relating to the relevant phenomenon is large (compare the two distributions in Figure 3.2).

In summary, our results regarding close replications show that low heterogeneity can be achieved in psychology. However, in meta-analyses (conceptual replications), it is typically large, and we did not find evidence to clarify the possible reasons for this.

## 7.2 Study 1b: Influence of bias: Questionable research practices and publication bias

Extensive research has found that both questionable research practices (QRPs) and publication bias are likely to cause biases in reported results (Ferguson & Brannick, 2012; John, Loewenstein, & Prelec, 2012; Kühberger, Fritz, & Scherndl, 2014; LeBel & Peters, 2011; Levine, Asada, & Carpenter, 2009; Sterling, 1959). In Study 1b, we used computer simulations to model the impact that these influences might have on observed levels of heterogeneity, for the first time. These simulations suggest that one-tailed versus two-tailed testing has a strong impact on bias in heterogeneity estimates. One-tailed testing leads to underestimation of true heterogeneity under conditions of bias, however, under two-tailed testing, these biases inflate true heterogeneity. Future research into heterogeneity should take this into account. In contrast to this, the influence of questionable research practices on observed levels of heterogeneity proved to be less important.

The levels of bias we found in Study 1b were low relative to the average observed heterogeneity in Study 1a, suggesting that there is no reason to believe that our conclusions regarding the large magnitude of heterogeneity and their implications for research practice in psychology are incorrect.

## 7.3 Study 2: The largest effects in psychology

Study 2 examined very-large-effects in psychology, from meta-analyses published in the last 5 years. We found that very-large-effects are more likely to be found in some sub-disciplines in

psychology than in others, with meta-analyses in organisational and cognitive psychology particularly more likely to find very-large-effects. We also found that a particularly high proportion of meta-analyses are conducted in clinical psychology, with several very-large-effects noted in this discipline as well.

Potential sources of error or bias (such as QRPs and publication) can inflate effect size, and this could mean that these errors are overrepresented in the very-large-effect meta-analyses. Most meta-analyses unfortunately do not assess or report bias of original studies. Assessing bias within the meta-analyses in our samples from reported data on this was therefore not feasible. However, we did find that our very-large-effects sample included multiple meta-analyses into the same effect, which might provide some assurance that these effects are not inflated due to error in a specific meta-analysis.

We found no difference between quality of meta-analyses in our very-large-effects sample and our control sample, and no indication that meta-analyses including only small number of studies (small  $k$ ) were overrepresented in our very-large-effect sample, providing tentative evidence that there is no strong role of bias in our sample. However, we are aware that we did not examine all sources of bias as this was outside the scope of this study.

#### **7.4 Study 3: Applying heterogeneity in a new context**

As we developed here, heterogeneity can be used to generate new predictions, or add a new perspective to existing theoretical debates. We examined this in Study 3, in the context of cross-cultural variability of psychological sex differences, in an attempt to provide a new perspective on the issue of whether sex differences may have evolved as inherited adaptations (the evolutionary perspective, for example, see Buss, 1989), or whether sex differences may reflect differences in the roles that are occupied by males and females in society (the social structuralist perspective, for example, see Eagly & Wood, 1999).

We identified key flaws in the central arguments about the origin of sex differences, and determined that heterogeneity could make a valuable contribution to evaluating each perspective. It is clear that the sex differences in our target sample do not look alike when viewed from this new perspective. For example, we proposed that our results regarding the low heterogeneity and *Flop Index* for the mental rotation sex differences in our sample suggested this was more likely to have biological underpinnings than sex differences in mate preferences in ambition. However, we acknowledge the limitations of Study 3 whereby our



biological comparison data (height and 2d:4d digit ratio) showed greater heterogeneity than expected. We are cautiously optimistic that the addition of more comparison data in future might prove fruitful in exploring the application of heterogeneity to the origins of sex differences further.

#### 7.4.1 Unreliable sources of measurement

One possible source of heterogeneity identified in study 3 was that introduced by unreliable sources of measurement, for example use of unvalidated questionnaires. Although we were not able to measure and assess this directly, we found that levels of heterogeneity were higher in studies conducted across multiple countries using unvalidated measures, compared to those using highly reliable and valid measures. This initial finding could perhaps be usefully explored further in future.

### 7.5 Conclusions

Both study 1a and study 2 took a descriptive approach to examining heterogeneity and the largest effects in psychology, respectively. As mentioned earlier, descriptive work has benefit for providing an overview of research and bringing multiple meta-analyses together, and this kind of work has been beneficial in other scientific disciplines from chemistry (see section 5.4.6) to geosciences (see section 1.4). This can begin to illuminate patterns or raise questions for subsequent investigations of the data. Study 1a provided an estimate of heterogeneity across multiple meta-analyses from a broader sample than previously conducted estimates for either close replications or conceptual replications (for example, see Stanley et al., 2018; Van Erp et al., 2017). Our examination of heterogeneity shows that focus on effect size alone can provide a skewed perspective on what is known about a particular effect, and highlights the importance of also considering levels of heterogeneity and the *Flop Index* to consider consistency of an effect. In the current project, the descriptive work undertaken in Study 1a also provided us with a new tool for theory testing, which we subsequently applied in Study 3.

## Appendices

### Appendix A

Study 1a: References for all included meta-analyses and close replications

#### *Close replications*

- Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R., Aykutoğlu, B., Bahník, Š., ... Caprariello, P. (2016). Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, 11(5), 750-764.
- Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... Boucher, L. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68-82.
- Eerland, A., Sherrill, A., Magliano, J., Zwaan, R., Arnal, J., Aucoin, P., ... Carlucci, M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, 11(1), 158-171.
- Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., ... Bruyneel, S. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546-573.
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... Brumbaugh, C. C. (2014). Investigating variation in replicability. *Social Psychology*, 45(3), 142-152.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., . . . Bahník, Š. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490.
- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. ... Blouin-Hudon, E.-M. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917-928.

#### *Meta-analyses*

#### Cognitive

- Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*(2), 207-245.
- Allen, M., Berkowitz, S., Hunt, S., & Louden, A. (1999). A meta-analysis of the impact of forensics and communication education on critical thinking. *Communication Education, 48*(1), 18-30.
- Alvarez, J. A., & Emory, E. (2006). Executive function and the frontal lobes: A meta-analytic review. *Neuropsychology Review, 16*(1), 17-42.
- Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuhl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. *Psychonomic Bulletin and Review, 22*(2), 366-377.
- Baker, C. A., Peterson, E., Pulos, S., & Kirkland, R. A. (2014). Eyes and IQ: A meta-analysis of the relationship between intelligence and "Reading the Mind in the Eyes". *Intelligence, 44*, 78-92.
- Beaujean, A. A. (2005). Heritability of cognitive abilities as measured by mental chronometric tasks: A meta-analysis. *Intelligence, 33*(2), 187-201.
- Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin and Review, 21*(1), 149-154.
- Chen, Q., & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. *Acta Psychologica, 148*, 163-172.
- D'Alessio, D., & Allen, M. (2002). Selective exposure and dissonance after decisions. *Psychological Reports, 91*(2), 527-532.
- Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review, 3*(4), 422-433.
- Dato-on, M. C., & Dahlstrom, R. (2003). A meta-analytic investigation of contrast effects in decision making. *Psychology and Marketing, 20*(8), 707-731.
- Dewald, J. F., Meijer, A. M., Oort, F. J., Kerkhof, G. A., & Bögels, S. M. (2010). The influence of sleep quality, sleep duration and sleepiness on school performance in children and adolescents: A meta-analytic review. *Sleep Medicine Reviews, 14*(3), 179-189.
- Doyle, R. A., & Voyer, D. (2016). Stereotype manipulation effects on math and spatial test performance: A meta-analysis. *Learning and Individual Differences, 47*, 103-116.
- Eastvold, A. D., Belanger, H. G., & Vanderploeg, R. D. (2012). Does a third party observer affect neuropsychological test performance? It depends. *The Clinical Neuropsychologist, 26*(3), 520-541.
- Evers, E. R., & Lakens, D. (2014). Revisiting Tversky's diagnosticity principle. *Frontiers in Psychology, 5*, 875.

- Fawcett, J. M. (2013). The production effect benefits performance in between-subject designs: A meta-analysis. *Acta Psychologica*, 142(1), 1-5.
- Herlitz, A., & Loven, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analysis review. *Visual Cognition*, 21(9-10), 1306-1336.
- Hinshaw, K. E. (1991). The effects of mental practice on motor skill performance: Critical evaluation and meta-analysis. *Imagination, Cognition and Personality*, 11(1), 3-35.
- Huff, M. J., Bodner, G. E., & Fawcett, J. M. (2015). Effects of distinctive encoding on correct and false memory: A meta-analytic review of costs and benefits and their origins in the DRM paradigm. *Psychonomic Bulletin and Review*, 22(2), 349-365.
- Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin and Review*, 21(4), 861-883.
- Lucas, M. (1999). Context effects in lexical access: A meta-analysis. *Memory and Cognition*, 27(3), 385-398.
- Maeda, Y., & Yoon, S. Y. (2013). A meta-analysis on gender differences in mental rotation ability measured by the Purdue spatial visualization tests: Visualization of rotations (PSVT: R). *Educational Psychology Review*, 25(1), 69-94.
- McDaniel, M. A. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence*, 33(4), 337-346.
- McMorris, T., & Hale, B. J. (2012). Differential effects of differing intensities of acute exercise on speed and accuracy of cognition: A meta-analytical investigation. *Brain and Cognition*, 80(3), 338-351.
- Meijer, E. H., Selle, N. K., Elber, L., & Ben-Shakhar, G. (2014). Memory detection with the Concealed Information Test: A meta analysis of skin conductance, respiration, heart rate, and P300 data. *Psychophysiology*, 51(9), 879-904.
- Merikle, P. M., & Daneman, M. (1996). Memory for unconsciously perceived events: Evidence from anesthetized patients. *Consciousness and Cognition*, 5(4), 525-541.
- Moxley, J. H., & Charness, N. (2013). Meta-analysis of age and skill effects on recalling chess positions and selecting the best move. *Psychonomic Bulletin and Review*, 20(5), 1017-1022.
- Nieuwenstein, M. R., Wierenga, T., Morey, R. D., Wicherts, J. M., Blom, T. N., Wagenmakers, E.-J., & van Rijn, H. (2015). On making the right choice: A meta-analysis and large-scale replication attempt of the unconscious thought advantage. *Judgment and Decision Making*, 10(1), 1-17.
- Park, J. (2005). Effect of arousal and retention delay on memory: A meta-analysis. *Psychological Reports*, 97(2), 339-355.

- Pässler, K., Beinicke, A., & Hell, B. (2015). Interests and intelligence: A meta-analysis. *Intelligence*, 50, 30-51.
- Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Mozart effect-Shmozart effect: A meta-analysis. *Intelligence*, 38(3), 314-323.
- Powers, K. L., Brooks, P. J., Aldrich, N. J., Palladino, M. A., & Alfieri, L. (2013). Effects of video-game play on information processing: A meta-analytic investigation. *Psychonomic Bulletin and Review*, 20(6), 1055-1079.
- Preiss, R. W., Wheelless, L. R., & Allen, M. (1990). Potential cognitive processes and consequences of receiver apprehension: A meta-analytic review. *Journal of Social Behavior and Personality*, 5(2), 155-172.
- Redick, T. S., & Lindsey, D. R. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin and Review*, 20(6), 1102-1113.
- Reinwein, J. (2012). Does the modality effect exist? And if so, which modality effect? *Journal of Psycholinguistic Research*, 41(1), 1-32.
- Rice, D. R., & Mullen, B. (2003). Isaac, Ishmael, and Janus: Past and future lessons regarding the ethnic categorization of faces. *Applied Cognitive Psychology*, 17(9), 1129-1147.
- Rojahn, K., & Pettigrew, T. F. (1992). Memory for schema-relevant information: A meta-analytic resolution. *British Journal of Social Psychology*, 31(2), 81-109.
- Schaubroeck, J., & Muralidhar, K. (1991). A meta-analysis of the relative effects of tabular and graphic display formats on decision-making performance. *Human Performance*, 4(2), 127-145.
- Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist*, 50(2), 138-166.
- Silverman, I. W. (2003). Gender differences in delay of gratification: A meta-analysis. *Sex Roles*, 49(9-10), 451-463.
- Sommer, I. E., Aleman, A., Bouma, A., & Kahn, R. S. (2004). Do women really have more bilateral language representation than men? A meta-analysis of functional imaging studies. *Brain*, 127(8), 1845-1852.
- Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence*, 53, 92-101.
- te Nijenhuis, J., Jongeneel-Grimen, B., & Kirkegaard, E. O. (2014). Are Headstart gains on the g factor? A meta-analysis. *Intelligence*, 46, 209-215.
- te Nijenhuis, J., van den Hoek, M., & Armstrong, E. L. (2015). Spearman's hypothesis and Amerindians: A meta-analysis. *Intelligence*, 50, 87-92.

- Tybur, J. M., Laakasuo, M., Ruff, J., & Klauke, F. (2016). How pathogen cues shape impressions of foods: The omnivore's dilemma and functionally specialized conditioning. *Evolution and Human Behavior*, 37, 376-386.
- Voss, M. W., Kramer, A. F., Basak, C., Prakash, R. S., & Roberts, B. (2010). Are expert athletes 'expert' in the cognitive laboratory? A meta-analytic review of cognition and sport expertise. *Applied Cognitive Psychology*, 24(6), 812-826.
- Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: A meta-analysis. *Psychonomic Bulletin and Review*, 18(2), 267-277.
- Voyer, D., & Jansen, P. (2017). Motor expertise and performance in spatial tasks: A meta-analysis. *Human Movement Science*, 54, 110-124.
- Voyer, D., Postma, A., Brake, B., & Imperato-McGinley, J. (2007). Gender differences in object location memory: A meta-analysis. *Psychonomic Bulletin and Review*, 14(1), 23-38.
- Wang, P., Liu, H.-H., Zhu, X.-T., Meng, T., Li, H.-J., & Zuo, X.-N. (2016). Action video game training for healthy adults: A meta-analytic study. *Frontiers in Psychology*, 7, 907.

#### Organisational

- Banks, G. C., Batchelor, J. H., Seers, A., O'Boyle, E. H., Pollack, J. M., & Gower, K. (2014). What does team-member exchange bring to the party? A meta-analytic review of team and leader social exchange. *Journal of Organizational Behavior*, 35(2), 273-295.
- Banks, G. C., McCauley, K. D., Gardner, W. L., & Guler, C. E. (2016). A meta-analytic review of authentic and transformational leadership: A test for redundancy. *The Leadership Quarterly*, 27(4), 634-652.
- Beal, D. J., Cohen, R. R., Burke, M. J., & McLendon, C. L. (2003). Cohesion and performance in groups: A meta-analytic clarification of construct relations. *Journal of Applied Psychology*, 88(6), 989-1004.
- Berry, C. M., Sackett, P. R., & Landers, R. N. (2007). Revisiting interview-cognitive ability relationships: Attending to specific range restriction mechanisms in meta-analysis. *Personnel Psychology*, 60(4), 837-874.
- Beus, J. M., & Whitman, D. S. (2012). The relationship between typical and maximum performance: A meta-analytic examination. *Human Performance*, 25(5), 355-376.
- Boer, D., Deinert, A., Homan, A. C., & Voelpel, S. C. (2016). Revisiting the mediating role of leader-member exchange in transformational leadership: The differential impact model. *European Journal of Work and Organizational Psychology*, 25(6), 883-899.
- Bowling, N. A., Khazon, S., Meyer, R. D., & Burrus, C. J. (2015). Situational strength as a moderator of the relationship between job satisfaction and job performance: A meta-analytic examination. *Journal of Business and Psychology*, 30(1), 89-104.

- Burke, M. J., Salvador, R. O., Smith-Crowe, K., Chan-Serafin, S., Smith, A., & Sonesh, S. (2011). The dread factor: How hazards and safety training influence learning and performance. *Journal of Applied Psychology, 96*(1), 46-70.
- Butts, M. M., Casper, W. J., & Yang, T. S. (2013). How important are work-family support policies? A meta-analytic investigation of their effects on employee outcomes. *Journal of Applied Psychology, 98*(1), 1-25.
- Choi, D., Oh, I.-S., & Colbert, A. E. (2015). Understanding organizational commitment: A meta-analytic examination of the roles of the five-factor model of personality and culture. *Journal of Applied Psychology, 100*(5), 1542-1567.
- Cohen, A., & Hudecek, N. (1993). Organizational commitment-turnover relationship across occupational groups: A meta-analysis. *Group and Organization Management, 18*(2), 188-213.
- Costanza, D. P., Badger, J. M., Fraser, R. L., Severt, J. B., & Gade, P. A. (2012). Generational differences in work-related attitudes: A meta-analysis. *Journal of Business and Psychology, 27*(4), 375-394.
- de Wit, F. R., Greer, L. L., & Jehn, K. A. (2012). The paradox of intragroup conflict: A meta-analysis. *Journal of Applied Psychology, 97*(2), 360-390.
- Donovan, J. J., & Radosevich, D. J. (1998). The moderating role of goal commitment on the goal difficulty-performance relationship. *Journal of Applied Psychology, 83*(2), 308-315.
- Driskell, J. E., Willis, R. P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77*(5), 615-622.
- Ernst Kossek, E., & Ozeki, C. (1998). Work-family conflict, policies, and the job-life satisfaction relationship: A review and directions for organizational behavior-human resources research. *Journal of Applied Psychology, 83*(2), 139-149.
- Foels, R., Driskell, J. E., Mullen, B., & Salas, E. (2000). The effects of democratic leadership on group member satisfaction and integration. *Small Group Research, 31*(6), 676-701.
- Gonzalez-Mulé, E., Mount, M. K., & Oh, I.-S. (2014). A meta-analysis of the relationship between general mental ability and nontask performance. *Journal of Applied Psychology, 99*(6), 1222-1243.
- Grijalva, E., & Newman, D. A. (2015). Narcissism and counterproductive work behavior (CWB): Meta-analysis and consideration of collectivist culture, Big Five personality, and narcissism's facet structure. *Applied Psychology, 64*(1), 93-126.
- Harari, M. B., Reaves, A. C., & Viswesvaran, C. (2016). Creative and innovative performance: A meta-analysis of relationships with task, citizenship, and counterproductive job performance dimensions. *European Journal of Work and Organizational Psychology, 25*(4), 495-511.

- Harari, M. B., & Rudolph, C. W. (2017). The effect of rater accountability on performance ratings: A meta-analytic review. *Human Resource Management Review*, 27(1), 121-133.
- Harms, P., Credé, M., Tynan, M., Leon, M., & Jeung, W. (2016). Leadership and stress: A meta-analytic review. *The Leadership Quarterly*.
- Jackson, S. E., & Schuler, R. S. (1985). A meta-analysis and conceptual critique of research on role ambiguity and role conflict in work settings. *Organizational Behavior and Human Decision Processes*, 36(1), 16-78.
- Jiang, K., Liu, D., McKay, P. F., Lee, T. W., & Mitchell, T. R. (2012). When and how is job embeddedness predictive of turnover? A meta-analytic investigation. *Journal of Applied Psychology*, 97(5), 1077-1096.
- Jin, J., & Rounds, J. (2012). Stability and change in work values: A meta-analysis of longitudinal studies. *Journal of Vocational Behavior*, 80(2), 326-339.
- Joseph, D. L., Dhanani, L. Y., Shen, W., McHugh, B. C., & McCord, M. A. (2015). Is a happy leader a good leader? A meta-analytic investigation of leader trait affect and leadership. *The Leadership Quarterly*, 26(4), 557-576.
- Keim, A. C., Landis, R. S., Pierce, C. A., & Earnest, D. R. (2014). Why do employees worry about their jobs? A meta-analytic review of predictors of job insecurity. *Journal of Occupational Health Psychology*, 19(3), 269-290.
- Klein, C., DiazGranados, D., Salas, E., Le, H., Burke, C. S., Lyons, R., & Goodwin, G. F. (2009). Does team building work? *Small Group Research*, 40(2), 181-222.
- Knight, A. P., & Eisenkraft, N. (2015). Positive is usually good, negative is not always bad: The effects of group affect on social integration and task performance. *Journal of Applied Psychology*, 100(4), 1214-1227.
- Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, 100(1), 128-161.
- Koh, C. W., Shen, W., & Lee, T. (2016). Black-White mean differences in job satisfaction: A meta-analysis. *Journal of Vocational Behavior*, 94, 131-143.
- Liu, D., Jiang, K., Shalley, C. E., Keem, S., & Zhou, J. (2016). Motivational mechanisms of employee creativity: A meta-analytic examination and theoretical extension of the creativity literature. *Organizational Behavior and Human Decision Processes*, 137, 236-263.
- Liu, Y., Guo, J., & Chi, N. (2015). The antecedents and performance consequences of proactive environmental strategy: A meta-analytic review of national contingency. *Management and Organization Review*, 11(03), 521-557.



- Martocchio, J. J., & O'Leary, A. M. (1989). Sex differences in occupational stress: A meta-analytic review. *Journal of Applied Psychology, 74*(3), 495-501.
- Murphy, G. C., & Athanasou, J. A. (1999). The effect of unemployment on mental health. *Journal of Occupational and Organizational Psychology, 72*(1), 83-99.
- Nicolaides, V. C., LaPort, K. A., Chen, T. R., Tomassetti, A. J., Weis, E. J., Zaccaro, S. J., & Cortina, J. M. (2014). The shared leadership of teams: A meta-analysis of proximal, distal, and moderating relationships. *The Leadership Quarterly, 25*(5), 923-942.
- Orlitzky, M., & Hirokawa, R. Y. (2001). To err is human, to correct for it divine a meta-analysis of research testing the functional theory of group decision-making effectiveness. *Small Group Research, 32*(3), 313-341.
- Park, H. I., Jacob, A. C., Wagner, S. H., & Baiden, M. (2014). Job control and burnout: A meta-analytic test of the Conservation of Resources model. *Applied Psychology, 63*(4), 607-642.
- Parks, K. M., & Steelman, L. A. (2008). Organizational wellness programs: A meta-analysis. *Journal of Occupational Health Psychology, 13*(1), 58-68.
- Rauch, A., Rosenbusch, N., Unger, J., & Frese, M. (2016). The effectiveness of cohesive and diversified networks: A meta-analysis. *Journal of Business Research, 69*(2), 554-568.
- Robertson, I. T. (1989). Work-sample tests of trainability: A meta-analysis. *Journal of Applied Psychology, 74*(3), 402-410.
- Rockstuhl, T., Dulebohn, J. H., Ang, S., & Shore, L. M. (2012). Leader-member exchange (LMX) and culture: A meta-analysis of correlates of LMX across 23 countries. *Journal of Applied Psychology, 97*(6), 1097-1130.
- Sadri, G., & Robertson, I. T. (1993). Self-efficacy and work-related behaviour: A review and meta-analysis. *Applied Psychology, 42*(2), 139-152.
- Schmidt, S., Roesler, U., Kusserow, T., & Rau, R. (2014). Uncertainty in the workplace: Examining role ambiguity and role conflict, and their link to depression – a meta-analysis. *European Journal of Work and Organizational Psychology, 23*(1), 91-106.
- Theeboom, T., Beersma, B., & van Vianen, A. E. (2013). Does coaching work? A meta-analysis on the effects of coaching on individual level outcomes in an organizational context. *The Journal of Positive Psychology, 9*(1), 1-18.
- Thorsteinson, T. J. (2003). Job attitudes of part-time vs. full-time workers: A meta-analytic review. *Journal of Occupational and Organizational Psychology, 76*(2), 151-177.
- Vaden, E. A., & Hall, S. (2005). The effect of simulator platform motion on pilot training transfer: A meta-analysis. *The International Journal of Aviation Psychology, 15*(4), 375-393.

- Verquer, M. L., Beehr, T. A., & Wagner, S. H. (2003). A meta-analysis of relations between person-organization fit and work attitudes. *Journal of Vocational Behavior*, 63(3), 473-489.
- Whelpley, C. E., & McDaniel, M. A. (2016). Self-esteem and counterproductive work behaviors: A systematic review. *Journal of Managerial Psychology*, 31(4), 850-863.
- Wiersma, U. J. (1992). The effects of extrinsic rewards in intrinsic motivation: A meta-analysis. *Journal of Occupational and Organizational Psychology*, 65(2), 101-114.

### Social

- Anderson, K. J., & Leaper, C. (1998). Meta-analyses of gender effects on conversational interruption: Who, what, when, where, and how. *Sex Roles*, 39(3), 225-252.
- Anestis, M. D., Soberay, K. A., Gutierrez, P. M., Hernández, T. D., & Joiner, T. E. (2014). Reconsidering the link between impulsivity and suicidal behavior. *Personality and Social Psychology Review*, 18(4), 366-386.
- Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, 77(3), 281-311.
- Batalha, L., Reynolds, K. J., & Newbiggin, C. A. (2011). All else being equal: Are men always higher in social dominance orientation than women? *European Journal of Social Psychology*, 41(6), 796-806.
- Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, 41(4), 540-558.
- Blondé, J., & Girandola, F. (2016). Revealing the elusive effects of vividness: A meta-analysis of empirical evidences assessing the effect of vividness on persuasion. *Social Influence*, 11(2), 111-129.
- Bornstein, R. F., & Cecero, J. J. (2000). Deconstructing dependency in a five-factor world: A meta-analytic review. *Journal of Personality Assessment*, 74(2), 324-343.
- Braithwaite, R., & Corr, P. J. (2016). Hans Eysenck, education and the experimental approach: A meta-analysis of academic capabilities in university students. *Personality and Individual Differences*, 103, 163-171.
- Byron, K., & Khazanchi, S. (2011). A meta-analytic investigation of the relationship of state and trait anxiety to performance on figural and verbal creative tasks. *Personality and Social Psychology Bulletin*, 37(2), 269-283.
- Calin-Jageman, R. J., & Caldwell, T. L. (2014). Replication of the superstition and performance study by Damisch, Stoberock, and Mussweiler (2010). *Social Psychology*, 45(3), 239-245.

- Costello, K., & Hodson, G. (2014). Lay beliefs about the causes of and solutions to dehumanization and prejudice: Do nonexperts recognize the role of human-animal relations? *Journal of Applied Social Psychology, 44*(4), 278-288.
- Craddock, E., van Dellen, M. R., Novak, S. A., & Ranby, K. W. (2015). Influence in relationships: A meta-analysis on health-related social control. *Basic and Applied Social Psychology, 37*(2), 118-130.
- DeCoster, J., & Claypool, H. M. (2004). A meta-analysis of priming effects on impression formation supporting a general model of informational biases. *Personality and Social Psychology Review, 8*(1), 2-27.
- Deshpande, S. P., & Viswesvaran, C. (1992). Is cross-cultural training of expatriate managers effective: A meta analysis. *International Journal of Intercultural Relations, 16*(3), 295-310.
- Devine, D. J., & Philips, J. L. (2001). Do smarter teams do better a meta-analysis of cognitive ability and team performance. *Small Group Research, 32*(5), 507-532.
- Donahue, M. J. (1985). Intrinsic and extrinsic religiousness: Review and meta-analysis. *Journal of Personality and Social Psychology, 48*(2), 400-419.
- Drescher, A., & Schultheiss, O. C. (2016). Meta-analytic evidence for higher implicit affiliation and intimacy motivation scores in women, compared to men. *Journal of Research in Personality, 64*, 1-10.
- Edens, J. F., Campbell, J. S., & Weir, J. M. (2007). Youth psychopathy and criminal recidivism: A meta-analysis of the psychopathy checklist measures. *Law and Human Behavior, 31*(1), 53-75.
- Georgesesen, J. C., & Harris, M. J. (1998). Why's my boss always holding me down? A meta-analysis of power effects on performance evaluations. *Personality and Social Psychology Review, 2*(3), 184-195.
- Gerber, J., & Wheeler, L. (2009). On being rejected: A meta-analysis of experimental research on rejection. *Perspectives on Psychological Science, 4*(5), 468-488.
- Goemans, A., van Geel, M., & Vedder, P. (2015). Over three decades of longitudinal research on the development of foster children: A meta-analysis. *Child Abuse and Neglect, 42*, 121-134.
- Gully, S. M., Devine, D. J., & Whitney, D. J. (2012). A meta-analysis of cohesion and performance: Effects of level of analysis and task interdependence. *Small Group Research, 43*(6), 702-725.
- Hauch, V., Sporer, S. L., Michael, S. W., & Meissner, C. A. (2016). Does training improve the detection of deception? A meta-analysis. *Communication Research, 43*(3), 283-343.

- Hu, T., Zhang, D., & Wang, J. (2015). A meta-analysis of the trait resilience and mental health. *Personality and Individual Differences, 76*, 18-27.
- IJzerman, H., Blanken, I., Brandt, M. J., Oerlemans, J., Van den Hoogenhof, M. M., Franken, S. J., & Oerlemans, M. W. (2014). Sex differences in distress from infidelity in early adulthood and in later life. *Social Psychology, 45*(3), 202-208.
- Leach, C. W., & Cidam, A. (2015). When is shame linked to constructive approach orientation? A meta-analysis. *Journal of Personality and Social Psychology, 109*(6), 983-1002.
- Lee, S., Moon, S.-I., & Feeley, T. H. (2016). A meta-analytic review of the legitimization of paltry favors compliance strategy. *Psychological Reports, 118*(3), 748-771.
- Littleton, H. L., Bye, K., Buck, K., & Amacker, A. (2010). Psychosocial stress during pregnancy and perinatal outcomes: A meta-analytic review. *Journal of Psychosomatic Obstetrics and Gynecology, 31*(4), 219-228.
- Masi, C. M., Chen, H.-Y., Hawkley, L. C., & Cacioppo, J. T. (2010). A meta-analysis of interventions to reduce loneliness. *Personality and Social Psychology Review, 15*(3), 219-266.
- McMahan, E. A., & Estes, D. (2015). The effect of contact with natural environments on positive and negative affect: A meta-analysis. *The Journal of Positive Psychology, 10*(6), 507-519.
- Morgan, R. D., Flora, D. B., Kroner, D. G., Mills, J. F., Varghese, F., & Steffan, J. S. (2012). Treating offenders with mental illness: A research synthesis. *Law and Human Behavior, 36*(1), 37-50.
- Mullen, B., & Hu, L. t. (1988). Social projection as a function of cognitive mechanisms: Two meta-analytic integrations. *British Journal of Social Psychology, 27*(4), 333-356.
- Nadler, J. T., & Clark, M. (2011). Stereotype threat: A meta-analysis comparing African Americans to Hispanic Americans. *Journal of Applied Social Psychology, 41*(4), 872-890.
- O'Connor, W. E., Morrison, T. G., McLeod, L. D., & Anderson, D. (1996). A meta-analytic review of the relationship between gender and belief in a just world. *Journal of Social Behavior and Personality, 11*(1), 141-148.
- Payne, K. T., & Marcus, D. K. (2008). The efficacy of group psychotherapy for older adult clients: A meta-analysis. *Group Dynamics: Theory, Research and Practice, 12*(4), 268-278.
- Randall, D. M., & Wolff, J. A. (1994). The time interval in the intention-behaviour relationship: Meta-analysis. *British Journal of Social Psychology, 33*(4), 405-418.
- Rise, J., Sheeran, P., & Hukkelberg, S. (2010). The role of self-identity in the Theory of Planned Behavior: A meta-analysis. *Journal of Applied Social Psychology, 40*(5), 1085-1105.

- Rivis, A., Sheeran, P., & Armitage, C. J. (2009). Expanding the affective and normative components of the Theory of Planned Behavior: A meta-analysis of anticipated affect and moral norms. *Journal of Applied Social Psychology, 39*(12), 2985-3019.
- Roccato, M., & Ricolfi, L. (2005). On the correlation between right-wing authoritarianism and social dominance orientation. *Basic and Applied Social Psychology, 27*(3), 187-200.
- Roisman, G. I., Holland, A., Fortuna, K., Fraley, R. C., Clausell, E., & Clarke, A. (2007). The Adult Attachment Interview and self-reports of attachment style: An empirical rapprochement. *Journal of Personality and Social Psychology, 92*(4), 678-697.
- Schütz, H., & Six, B. (1996). How strong is the relationship between prejudice and discrimination? A meta-analytic answer. *International Journal of Intercultural Relations, 20*(3), 441-462.
- Sibley, C. G., Osborne, D., & Duckitt, J. (2012). Personality and political orientation: Meta-analysis and test of a Threat-Constraint Model. *Journal of Research in Personality, 46*(6), 664-677.
- Stebly, N. M., Besirevic, J., Fulero, S. M., & Jimenez-Lorente, B. (1999). The effects of pretrial publicity on juror verdicts: A meta-analytic review. *Law and Human Behavior, 23*(2), 219-235.
- Terrizzi, J. A., Shook, N. J., & McDaniel, M. A. (2013). The behavioral immune system and social conservatism: A meta-analysis. *Evolution and Human Behavior, 34*(2), 99-108.
- van Osch, Y., Blanken, I., Meijs, M. H., & van Wolferen, J. (2015). A group's physical attractiveness is greater than the average attractiveness of its members: The group attractiveness effect. *Personality and Social Psychology Bulletin, 41*(4), 559-574.
- Vazire, S., & Funder, D. C. (2006). Impulsivity and the self-defeating behavior of narcissists. *Personality and Social Psychology Review, 10*(2), 154-165.
- Vedel, A. (2014). The Big Five and tertiary academic performance: A systematic review and meta-analysis. *Personality and Individual Differences, 71*, 66-76.
- Vicaria, I. M., & Dickens, L. (2016). Meta-analyses of the intra-and interpersonal outcomes of interpersonal coordination. *Journal of Nonverbal Behavior, 40*(4), 335-361.
- Zuckerman, M., Li, C., & Hall, J. A. (2016). When men and women differ in self-esteem and when they don't: A meta-analysis. *Journal of Research in Personality, 64*, 34-51.

Study 1a: Results based on Der Simonian Laird Hunter-Schmidt and Paule-Mandel estimators

*Descriptive statistics*

Descriptive statistics for study 1a

Sub-discipline	No. meta- analyses	Mean <i>k</i> per meta- analysis	Mean <i>d</i> ( <i>SD</i> )			Mean <i>T</i> ( <i>SD</i> )			Mean <i>FI</i> ( <i>SD</i> )			Heterogeneity ratings (lowest heterogeneity – 1 or 2)		
			DL	HS	PM	DL	HS	PM	DL	HS	PM	DL	HS	PM
Close replications	57	35.2	0.309 (0.337)	0.309 (0.337)	0.309 (0.337)	0.094 (0.071)	0.087 (0.068)	0.103 (0.079)	0.006 (0.010)	0.005 (0.009)	0.006 (0.010)			
Social	50	35.7				0.311 (0.111)	0.295 (0.109)	0.335 (0.118)	0.065 (0.056)	0.052 (0.052)	0.082 (0.070)			
Cognitive	50	36.5				0.321 (0.126)	0.300 (0.131)	0.356 (0.154)	0.072 (0.077)	0.071 (0.079)	0.090 (0.092)			
Organisational	50	38.3				0.345 (0.098)	0.326 (0.094)	0.393 (0.117)	0.084 (0.066)	0.075 (0.062)	0.096 (0.072)			
<i>Total</i>	<i>150</i>	<i>36.9</i>	<i>0.473</i> <i>(0.207)</i>	<i>0.473</i> <i>(0.207)</i>	<i>0.478</i> <i>(0.209)</i>	<i>0.325</i> <i>(0.110)</i>	<i>0.306</i> <i>(0.108)</i>	<i>0.357</i> <i>(0.121)</i>	<i>0.075</i> <i>(0.071)</i>	<i>0.069</i> <i>(0.071)</i>	<i>0.090</i> <i>(0.078)</i>	<i>0.288</i> <i>(0.116)</i>	<i>0.271</i> <i>(0.119)</i>	<i>0.344</i> <i>(0.111)</i>

*Note.* All means are winsorized. Means for Flop Index are only for meta-analyses which are statistically significant ( $p < .05$ ).

No significant differences between any models for *d*

### *Robust t-tests for group differences*

#### **Robust independent t-tests for $T$ between the groups**

DL: Meta-analyses vs close replications –  $t(94.940) = 10.427, p < .001$  (2-tailed)

Cognitive vs organisational –  $t(54.901) = 0.739, p = .463$  (2-tailed), cognitive vs social –  $t(57.307) = 0.334, p = .370$  (1-tailed), organisational vs social –  $t(57.021) = 1.165, p = .125$  (1-tailed)

HS: Meta-analyses vs close replications –  $t(97.332) = 10.385, p < .001$  (2-tailed)

Cognitive vs organisational –  $t(52.806) = 0.674, p = .503$  (2-tailed), cognitive vs social –  $t(56.431) = 0.282, p = .390$  (1-tailed), organisational vs social –  $t(56.640) = 1.080, p = .143$  (1-tailed)

PM: Meta-analyses vs close replications –  $t(94.133) = 10.415, p < .001$  (2-tailed)

Cognitive vs organisational –  $t(54.124) = 0.688, p = .494$  (2-tailed), cognitive vs social –  $t(54.133) = 0.459, p = .324$  (1-tailed), organisational vs social –  $t(58.000) = 1.340, p = .093$  (1-tailed)

#### **Robust independent t-tests for Cohen's $d$ in meta-analyses vs close replications**

DL:  $t(44.383) = 2.960, p = .005$

HS:  $t(44.412) = 2.939, p = .005$

PM:  $t(44.581) = 2.997, p = .004$

### Winsorized correlations between parameters

Winsorized correlations between parameters for close replications and meta-analyses

Sub-discipline	<i>T</i> and <i>d</i>			<i>T</i> and <i>k</i>			<i>d</i> and <i>FI</i> (significant only)		
	DL	HS	PM	DL	HS	PM	DL	HS	PM
Close replications	$r = .475$ , $p < .001$	$r = .486$ , $p < .001$	$r = .481$ , $p < .001$	$r = .132$ , $p = .332$	$r = .155$ , $p = .253$	$r = .128$ , $p = .345$	$r = -.368$ , $p = .056$	$r = -.379$ , $p = .048$	$r = -.324$ , $p = .093$
Social	$r = .519$ , $p < .001$	$r = .498$ , $p < .001$	$r = .460$ , $p = .001$						
Cognitive	$r = .338$ , $p = .019$	$r = .338$ , $p = .019$	$r = .347$ , $p = .016$						
Organisational	$r = .621$ , $p < .001$	$r = .585$ , $p < .001$	$r = .484$ , $p < .001$						
Total sample	$r = .489$ , $p < .001$	$r = .465$ , $p < .001$	$r = .428$ , $p < .001$	$r = .229$ , $p = .005$	$r = .292$ , $p < .001$	$r = .275$ , $p = .001$	$r = -.558$ , $p < .001$	$r = -.562$ , $p < .001$	$r = -.489$ , $p < .001$

Winsorized correlations for total meta-analysis sample ( $k = 150$ )

<i>T</i> and publication date			<i>T</i> and median sample size			<i>T</i> and heterogeneity ratings		
DL	HS	PM	DL	HS	PM	DL	HS	PM
$r = -.189$ , $p = .021$	$r = -.174$ , $p = .034$	$r = -.189$ , $p = .021$	$r = -.174$ , $p = .034$	$r = -.176$ , $p = .032$	$r = -.102$ , $p = .216$	$r = .121$ , $p = .142$	$r = .123$ , $p = .136$	$r = .089$ , $p = .279$



### Moderator analyses

Winsorized means (and standard deviations) for  $T$  and  $d$  for meta-analyses included in moderator sample.  $k = 22$ , except for theory-based, where  $k = 10$

Moderator	$d$			$T$		
	DL	HS	PM	DL	HS	PM
Main meta-analysis	0.372 (0.090)	0.365 (0.089)	0.478 (0.169)	0.463 (0.195)	0.462 (0.195)	0.472 (0.203)
Strongest moderator	0.328 (0.106)	0.308 (0.103)	0.372 (0.150)	0.577 (0.211)	0.577 (0.209)	0.601 (0.239)
Theory-based moderator	0.374 (0.066)	0.335 (0.049)	0.446 (0.121)	0.554 (0.201)	0.552 (0.199)	0.572 (0.219)

### **Robust paired t-tests for differences between main meta-analysis and moderators**

DL  $d$ : main vs strongest mod:  $t(13) = 2.187, p = .048$ , main vs theory mod:  $t(5) = 0.678, p = .528$

DL  $T$ : main vs strongest mod:  $t(13) = 1.393, p = .187$ , main vs theory mod:  $t(5) = 0.728, p = .499$

HS  $d$ : main vs strongest mod:  $t(13) = 2.213, p = .045$ , main vs theory mod:  $t(5) = 0.685, p = .524$

HS  $T$ : main vs strongest mod:  $t(13) = 2.050, p = .061$ , main vs theory mod:  $t(5) = 1.566, p = .178$

PM  $d$ : main vs strongest mod:  $t(13) = 1.991, p = .068$ , main vs theory mod:  $t(5) = 0.629, p = .557$

PM  $T$ : main vs strongest mod:  $t(13) = 2.000, p = .067$ , main vs theory mod:  $t(5) = 1.480, p = .199$

### Median date split

Winsorized means (and standard deviations) for  $T$  and  $d$  for meta-analyses included in median date split.

Median date split	$T$			$d$		
	DL	HS	PM	DL	HS	PM
Early dates	0.336 (0.097)	0.316 (0.097)	0.382 (0.107)	0.471 (0.158)	0.469 (0.159)	0.474 (0.159)

Late dates	0.342	0.328	0.390	0.446	0.446	0.449
	(0.110)	(0.104)	(0.141)	(0.175)	(0.176)	(0.177)

---

**Robust paired t-tests for differences between early and late dates in median date split**

DL ***d***:  $t(49) = 2.249, p = .029$

DL ***T***:  $t(49) = -1.007, p = .319$

HS ***d***:  $t(49) = 2.104, p = .040$

HS ***T***:  $t(49) = -1.238, p = .221$

PM ***d***:  $t(49) = 2.370, p = .022$

PM ***T***:  $t(49) = 0.291, p = .772$

*Predicting heterogeneity from effect size: Regression analyses*

**Regressions: predicting *T* from *d* using 150 meta-analyses (and using individual effect sizes from every study in each meta-analysis)**

DL: *d* significantly predicts *T*,  $R = .654, F(1,148) = 110.814, p < .001, T = 0.179 + 0.302d$  (*d* sig predicts *T*,  $R = .400, F(1,7224) = 1373.650, p < .001, T = 0.283 + 0.105d$ )

HS: *d* significantly predicts *T*,  $R = .620, F(1,148) = 92.444, p < .001, T = 0.171 + 0.278d$  (*d* sig predicts *T*,  $R = .391, F(1,7224) = 1301.422, p < .001, T = 0.273 + 0.101d$ )

PM: *d* significantly predicts *T*,  $R = .600, F(1,148) = 83.455, p < .001, T = 0.184 + 0.390d$  (*d* sig predicts *T*,  $R = .390, F(1,7224) = 1297.430, p < .001, T = 0.306 + 0.168d$ )

**Regressions: predicting *T* from *d* for 57 close replications**

DL: *d* significantly predicts *T*,  $R = .683, F(1,55) = 47.975, p < .001, T = 0.063 + 0.187d$

HS: *d* significantly predicts *T*,  $R = .682, F(1,55) = 47.702, p < .001, T = 0.057 + 0.183d$

PM: *d* significantly predicts *T*,  $R = .671, F(1,55) = 44.961, p < .001, T = 0.070 + 0.202d$

## Appendix B

### Study 1b: R code for simulations

```
#####
```

```
# housekeeping #
```

```
#####
```

```
# ---- remove all variables
```

```
rm(list=ls())
```

```
# ---- set working directory (Audrey's computer)
```

```
setwd("U:/My Documents/PhD/Heterogeneity study 1/R")
```

```
# ---- install various packages
```

```
# ---- verify the package is installed
```

```
# ---- load the library into R workspace
```

```
install.packages("xlsx")
```

```
any(grepl("xlsx",installed.packages()))
```

```
install.packages("e1071")
```

```
any(grepl("e1071",installed.packages()))
```

```
install.packages("Metrics")
```

```

any(grepl("Metrics",installed.packages()))
install.packages("metafor")
any(grepl("metafor",installed.packages()))
library("xlsx")
library("e1071")
library("Metrics")
library("metafor")

```

```

#####
#   d and MA   #
#####

```

```

# ---- myf.get.d()
# ---- takes experiment and returns unbiased estimate for Cohen's d
# ---- for unbiasing of d cf. Lakens 2013
myf.get.d <- function(experiment) {
  m_ctrlg <- mean(experiment[, 1], na.rm = TRUE)
  m_expg <- mean(experiment[, 2], na.rm = TRUE)
  var_ctrlg <- var(experiment[, 1], na.rm = TRUE)
  var_expg <- var(experiment[, 2], na.rm = TRUE)
  n <- nrow(experiment) - (sum(is.na(experiment)) / 2)

```

```

cohens_d <- (m_expg - m_ctrlg) / sqrt((var_ctrlg + var_expg) / 2)
d_unb <- cohens_d * (1 - 3 / (8 * n - 9))
return(d_unb)
}

```

```

# ---- myf.ma()
# ---- runs random-effects meta-analysis on k experiments, using the package Metafor
# ---- depending on MA_model, either DerSimonian-Laird, Hunter-Schmidt, or Paule-Mandel estimator is used
# ---- computes 95% CI around tau2
# ---- 'captured' indicates if CI did (1) or did not (0) capture true tau2
# ---- returns mean effect size, tau, captured, flop index, I2, and median N as 1x6 matrix
# ---- n in function argument is sample size per group

```

```

myf.ma <- function(d, n, MA_model) {
  n <- n * 2
  vi <- 4*(1+(d^2)/8)/n          # ---- computes variance (required by Metafor) as a vector, based on d and total n (assumes equal group sizes)
  res <- rma(d, vi, method=MA_model) # ---- Metafor function to run meta-analysis
  tau2 <- tau^2
  ma_tau2 <- res[[9]]
  ma_SEtau2 <- res[[10]]
  ll_tau2 <- ma_tau2 - (1.96 * ma_SEtau2)
  ul_tau2 <- ma_tau2 + (1.96 * ma_SEtau2)
}

```

```

if (ll_tau2[1] <= tau2 && ul_tau2 >= tau2) {
  captured <- 1
} else {
  captured <- 0
}

result <- matrix( , nrow = 1, ncol = 6)
result[1, 1] <- res[[1]]      # ---- d
result[1, 2] <- sqrt(res[[9]]) # ---- T (from T2)
result[1, 3] <- captured[1]
result[1, 4] <- (pnorm(0, mean = res[[1]], sd = sqrt(res[[9]])))[1] # ---- computation of flop index
result[1, 5] <- res[[23]]     # ---- l2
result[1, 6] <- median(n)     # ---- median sample size (across groups)
return(result)
}

#####
#   QRPs   the action happens in myf.aggr.qrps(), the last function in this #
#           section, which calls on the earlier functions in this section   #
#####

```

```

# ---- myf.unlist()
# ---- returns matrix from list
myf.unlist <- function(list, position, nrow, ncol) {
  v <- unlist(list[position])
  experiment <- matrix(v, nrow = nrow, ncol = ncol)
  return(experiment)
}

# ---- myf.alternative.dv()
# ---- creates experiment with two dependent variables (dv1 and dv2), which correlate r (set in parameters section)
# ---- each dv runs across 2 columns of a matrix; 1st col = control group, 2nd col = experimental group.
# ---- logic: true_dv (population variance = 1) is a background variable that is used to create dv1 and dv2
# ---- sd_noise reflects how much random noise needs to be added to true_dv for dv1 and dv2 to correlate r
# ---- dv1 is then created as true_dv + noise; consequently variance in dv1 > 1
# ---- dv1 is therefore divided by sd_noise to bring its (population) variance back to 1
# ---- same repeated for dv2
# ---- (effect_size + tau) is then added to scores to dv1 and dv2 in experimental group columns
myf.alternative.dv <- function(n, effect_size, tau, r) {
  true_dv <- matrix(rnorm(n * 2), nrow = n, ncol = 2)
  r <- sqrt(r)
  sd_noise <- sqrt((1 / r^2) - 1)

```

```

dv1 <- matrix(rnorm(n * 2, 0, sd_noise), nrow = n, ncol = 2)
dv1 <- (dv1 + true_dv) / sqrt(1 + sd_noise^2)
dv2 <- matrix(rnorm(n * 2, 0, sd_noise), nrow = n, ncol = 2)
dv2 <- (dv2 + true_dv) / sqrt(1 + sd_noise^2)
tau <- rnorm(1, 0, tau)
dv1[, 2] <- dv1[, 2] + effect_size + tau
dv2[, 2] <- dv2[, 2] + effect_size + tau
list_dvs <- list(dv1, dv2)
return(list_dvs)
}

```

```

# ---- creates moderator for experiment based on generating random binomial (0 or 1)
# ---- generates two copies of experiment, mod_0 removes scores corresponding to mod of 1
# ---- mod_1 removes scores corresponding to mod of 0
myf.moderators <- function (list_dvs, start_n) {
  dv1 <- myf.unlist(list_dvs, 1, start_n, 2)
  dv2 <- myf.unlist(list_dvs, 2, start_n, 2)
  mod <- rbinom((nrow(dv1)*2), 1, 0.5)
  dv1_mod0 <- dv1
  dv1_mod1 <- dv1
  for (i in 1:length(mod)) {

```



```
  if (mod[i] == 1) {  
    dv1_mod0[i] <- NA  
  }  
  if (mod[i] == 0) {  
    dv1_mod1[i] <- NA  
  }  
}  
dv2_mod0 <- dv2  
dv2_mod1 <- dv2  
for (i in 1:length(mod)) {  
  if (mod[i] == 1) {  
    dv2_mod0[i] <- NA  
  }  
  if (mod[i] == 0) {  
    dv2_mod1[i] <- NA  
  }  
}  
list_dvs_moderator <- list(dv1, dv1_mod0, dv1_mod1, dv2, dv2_mod0, dv2_mod1)  
return(list_dvs_moderator)  
}
```

```

# ---- myf.remove.outliers()
# ---- takes 6 datasets
# ---- makes copies and removes scores for which  $z > \text{abs}(z\_crit)$ 
# ---- returns 6 original datasets and 6 trimmed datasets (oToT...)
myf.remove.outliers <- function(list_dvs_moderator, n) {
  dv1 <- myf.unlist(list_dvs_moderator, 1, n, 2)
  dv1_mod0 <- myf.unlist(list_dvs_moderator, 2, n, 2)
  dv1_mod1 <- myf.unlist(list_dvs_moderator, 3, n, 2)
  dv2 <- myf.unlist(list_dvs_moderator, 4, n, 2)
  dv2_mod0 <- myf.unlist(list_dvs_moderator, 5, n, 2)
  dv2_mod1 <- myf.unlist(list_dvs_moderator, 6, n, 2)
  for (i in 1:6) {
    z_experiment <- switch(i, dv1, dv1_mod0, dv1_mod1, dv2, dv2_mod0, dv2_mod1)
    experiment_no_outliers <- z_experiment
    z_experiment[, 1] <- scale(z_experiment[, 1])
    z_experiment[, 2] <- scale(z_experiment[, 2])
    for(j in 1:nrow(z_experiment)) {
      if (is.na(z_experiment[j, 1]) == TRUE) {
      } else if (abs(z_experiment[j, 1]) > z_crit) {
        experiment_no_outliers[j, 1] = NA
      }
    }
  }
}

```

```

    if (is.na(z_experiment[j, 2]) == TRUE) {
    } else if (abs(z_experiment[j, 2]) > z_crit) {
      experiment_no_outliers[j, 2] = NA
    }
  }
  if (i == 1) {
    dv1_trim <- experiment_no_outliers
  } else if (i == 2) {
    dv1_mod0_trim <- experiment_no_outliers
  } else if (i == 3) {
    dv1_mod1_trim <- experiment_no_outliers
  } else if (i == 4) {
    dv2_trim <- experiment_no_outliers
  } else if (i == 5) {
    dv2_mod0_trim <- experiment_no_outliers
  } else {
    dv2_mod1_trim <- experiment_no_outliers
  }
}

list_dvs_moderator_outliers <- list(dv1, dv1_trim, dv1_mod0, dv1_mod0_trim, dv1_mod1, dv1_mod1_trim, dv2, dv2_trim, dv2_mod0, dv2_mod0_trim,
dv2_mod1, dv2_mod1_trim)

```

```

return(list_dvs_moderator_outliers)
}

# ---- myf.add.participants()
# ---- adds add_n new participants to old data sets
# ---- step 1: unlist original data sets
# ---- step 2: create list of new data sets (i.e. the new participants) and the moderator
# ---- step 3: unlist new data sets
# ---- step 4: stick old and new data sets together and return as list
myf.add.participants <- function(list_dvs_moderator_outliers, n, add_n, effect_size, tau, r) {
  dv1 <- myf.unlist(list_dvs_moderator_outliers, 1, n, 2)
  dv1_trim <- myf.unlist(list_dvs_moderator_outliers, 2, n, 2)
  dv1_mod0 <- myf.unlist(list_dvs_moderator_outliers, 3, n, 2)
  dv1_mod0_trim <- myf.unlist(list_dvs_moderator_outliers, 4, n, 2)
  dv1_mod1 <- myf.unlist(list_dvs_moderator_outliers, 5, n, 2)
  dv1_mod1_trim <- myf.unlist(list_dvs_moderator_outliers, 6, n, 2)
  dv2 <- myf.unlist(list_dvs_moderator_outliers, 7, n, 2)
  dv2_trim <- myf.unlist(list_dvs_moderator_outliers, 8, n, 2)
  dv2_mod0 <- myf.unlist(list_dvs_moderator_outliers, 9, n, 2)
  dv2_mod0_trim <- myf.unlist(list_dvs_moderator_outliers, 10, n, 2)
  dv2_mod1 <- myf.unlist(list_dvs_moderator_outliers, 11, n, 2)

```

```

dv2_mod1_trim <- myf.unlist(list_dvs_moderator_outliers, 12, n, 2)
list_new_subjects <- myf.alternative.dv(add_n, effect_size, tau, r)
list_new_moderator <- myf.moderators(list_new_subjects, add_n)
list_new_moderator_outliers <- myf.remove.outliers(list_new_moderator, add_n)
new_subjects_dv1 <- myf.unlist(list_new_moderator_outliers, 1, add_n, 2)
new_subjects_dv1_trim <- myf.unlist(list_new_moderator_outliers, 2, add_n, 2)
new_subjects_dv1_mod0 <- myf.unlist(list_new_moderator_outliers, 3, add_n, 2)
new_subjects_dv1_mod0_trim <- myf.unlist(list_new_moderator_outliers, 4, add_n, 2)
new_subjects_dv1_mod1 <- myf.unlist(list_new_moderator_outliers, 5, add_n, 2)
new_subjects_dv1_mod1_trim <- myf.unlist(list_new_moderator_outliers, 6, add_n, 2)
new_subjects_dv2 <- myf.unlist(list_new_moderator_outliers, 7, add_n, 2)
new_subjects_dv2_trim <- myf.unlist(list_new_moderator_outliers, 8, add_n, 2)
new_subjects_dv2_mod0 <- myf.unlist(list_new_moderator_outliers, 9, add_n, 2)
new_subjects_dv2_mod0_trim <- myf.unlist(list_new_moderator_outliers, 10, add_n, 2)
new_subjects_dv2_mod1 <- myf.unlist(list_new_moderator_outliers, 11, add_n, 2)
new_subjects_dv2_mod1_trim <- myf.unlist(list_new_moderator_outliers, 12, add_n, 2)
dv1 <- rbind(dv1, new_subjects_dv1)
dv1_trim <- rbind(dv1_trim, new_subjects_dv1_trim)
dv1_mod0 <- rbind(dv1_mod0, new_subjects_dv1_mod0)
dv1_mod0_trim <- rbind(dv1_mod0_trim, new_subjects_dv1_mod0_trim)
dv1_mod1 <- rbind(dv1_mod1, new_subjects_dv1_mod1)

```

```

dv1_mod1_trim <- rbind(dv1_mod1_trim, new_subjects_dv1_mod1_trim)
dv2 <- rbind(dv2, new_subjects_dv2)
dv2_trim <- rbind(dv2_trim, new_subjects_dv2_trim)
dv2_mod0 <- rbind(dv2_mod0, new_subjects_dv2_mod0)
dv2_mod0_trim <- rbind(dv2_mod0_trim, new_subjects_dv2_mod0_trim)
dv2_mod1 <- rbind(dv2_mod1, new_subjects_dv2_mod1)
dv2_mod1_trim <- rbind(dv2_mod1_trim, new_subjects_dv2_mod1_trim)
list_dvs_moderator_outliers <- list(dv1, dv1_trim, dv1_mod0, dv1_mod0_trim, dv1_mod1, dv1_mod1_trim, dv2, dv2_trim, dv2_mod0, dv2_mod0_trim,
dv2_mod1, dv2_mod1_trim)
  return(list_dvs_moderator_outliers)
}

```

165

```

# ---- myf.best.dataset()
# ---- receives 12 data sets with corresponding p-values
# ---- returns single data set with lowest p-value
myf.best.dataset <- function(list_dvs_moderator_outliers, n, p) {
  dataset <- which.min(p)
  dv1 <- myf.unlist(list_dvs_moderator_outliers, 1, n, 2)
  dv1_trim <- myf.unlist(list_dvs_moderator_outliers, 2, n, 2)
  dv1_mod0 <- myf.unlist(list_dvs_moderator_outliers, 3, n, 2)
  dv1_mod0_trim <- myf.unlist(list_dvs_moderator_outliers, 4, n, 2)

```

```
dv1_mod1 <- myf.unlist(list_dvs_moderator_outliers, 5, n, 2)
dv1_mod1_trim <- myf.unlist(list_dvs_moderator_outliers, 6, n, 2)
dv2 <- myf.unlist(list_dvs_moderator_outliers, 7, n, 2)
dv2_trim <- myf.unlist(list_dvs_moderator_outliers, 8, n, 2)
dv2_mod0 <- myf.unlist(list_dvs_moderator_outliers, 9, n, 2)
dv2_mod0_trim <- myf.unlist(list_dvs_moderator_outliers, 10, n, 2)
dv2_mod1 <- myf.unlist(list_dvs_moderator_outliers, 11, n, 2)
dv2_mod1_trim <- myf.unlist(list_dvs_moderator_outliers, 12, n, 2)
```

```
experiment <- switch(dataset,
```

```
  dv1,
  dv1_trim,
  dv1_mod0,
  dv1_mod0_trim,
  dv1_mod1,
  dv1_mod1_trim,
  dv2,
  dv2_trim,
  dv2_mod0,
  dv2_mod0_trim,
  dv2_mod1,
```

```

        dv2_mod1_trim)
    return(experiment)
}

# ---- myf.moderate.qrps() emulates use of 2 dvs and optional stopping during data collection
# ---- initial sample size drawn from distribution of empirically observed sample sizes from 150 meta-analyses
# ---- add_n and max_n determined; relevant for optional stopping of study (later in this function)
# ---- absolute_max_n is ceiling for max_n and specified in parameters section
# ---- calls for 2 dvs via myf.alternative.dv
# ---- keeps peeking until significant p-value or max_n reached; each time add_n new subjects added
# ---- for dv resulting in smaller p-value, it returns: effect size and n (for MA) and p-value (for publication bias)
myf.moderate.qrps <- function(absolute_max_n, effect_size, tau) {
  start_n <- n_vector[as.integer(runif(1, min=1, max= 7228))]
  if (start_n < 2) {
    start_n <- 2
  }
  add_n <- as.integer(start_n * 0.1)
  if (add_n < 1) {
    add_n <- 1
  }
  max_n <- 5 * start_n

```



```

if (max_n > absolute_max_n) {
  max_n <- absolute_max_n
}
list_dvs <- myf.alternative.dv(start_n, effect_size, tau, r)
current_n <- start_n
dv1 <- myf.unlist(list_dvs, 1, start_n, 2)
dv2 <- myf.unlist(list_dvs, 2, start_n, 2)
p1 <- myf.get.p(dv1)
p2 <- myf.get.p(dv2)
while (p1 > 0.05 && p2 > 0.05 && current_n < max_n) {
  list_new_subjects <- myf.alternative.dv(add_n, effect_size, tau, r)
  new_subjects1 <- myf.unlist(list_new_subjects, 1, add_n, 2)
  new_subjects2 <- myf.unlist(list_new_subjects, 2, add_n, 2)
  dv1 <- rbind(dv1, new_subjects1)
  dv2 <- rbind(dv2, new_subjects2)
  p1 <- myf.get.p(dv1)
  p2 <- myf.get.p(dv2)
  current_n <- current_n + add_n
}
if (p1 < p2) {
  d <- myf.get.d(dv1)

```

```

    p <- p1
  } else {
    d <- myf.get.d(dv2)
    p <- p2
  }
  results <- list(d, current_n, p)
  return(results)
}

```

```

# ---- myf.aggr.qrps() emulates use of 2 dvs, use of 2 level moderator, removal of outliers, and optional stopping during data collection,
# ---- initial sample size drawn from distribution of empirically observed sample sizes from 150 meta-analyses
# ---- add_n and max_n determined; relevant for optional stopping of study (later in this function)
# ---- absolute_max_n is ceiling for max_n and specified in parameters section
# ---- calls for 2 dvs via myf.alternative.dv
# ---- for each dv, creates 2 subsets based on moderator via myf.moderators
# ---- for each of now 6 datasets, creates additional dataset where outliers are removed via myf.remove.outliers
# ---- keeps peeking for 12 datasets until significant p-value or max_n reached; each time add_n new subjects added
# ---- for dataset resulting in smallest p-value, it returns: effect size and n (for MA) and p-value (for publication bias)
myf.aggr.qrps <- function(absolute_max_n, effect_size, tau) {
  start_n <- n_vector[as.integer(runif(1, min=1, max= 7228))]
  if (start_n < 2) {

```

```
start_n <- 2
}
add_n <- as.integer(start_n * 0.1)
if (add_n < 1) {
  add_n <- 1
}
max_n <- 5 * start_n
if (max_n > absolute_max_n) {
  max_n <- absolute_max_n
}
list_dvs <- myf.alternative.dv(start_n, effect_size, tau, r)
current_n <- start_n
list_dvs_moderator <- myf.moderators(list_dvs, start_n)
list_dvs_moderator_outliers <- myf.remove.outliers(list_dvs_moderator, start_n)
p <- myf.get.many_p(list_dvs_moderator_outliers, start_n)
while (min(p, na.rm = TRUE) > 0.05 && current_n < max_n) {
  list_dvs_moderator_outliers <- myf.add.participants(list_dvs_moderator_outliers, current_n, add_n, effect_size, tau, r)
  current_n <- current_n + add_n
  p <- myf.get.many_p(list_dvs_moderator_outliers, current_n)
}
experiment <- myf.best.dataset(list_dvs_moderator_outliers, current_n, p)
```

```

d <- myf.get.d(experiment)
p <- min(p, na.rm = TRUE)
results <- list(d, current_n - (sum(is.na(experiment)) / 2), p) # ---- n needs to be recalculated by counting number of empty data cells (e.g. removed outliers)
return(results)
}

```

```
#####
```

```
# p-values #
```

```
#####
```

```
# ---- myf.get.p()
```

```
# ---- p-value is based on one-tailed testing.
```

```
# ---- all effects in wrong direction (i.e. mean EG < mean CG) result in p = .999.
```

```
# ---- all effects in right direction (larger EG mean) results in one-tailed p-value.
```

```
# ---- returns p-value for t-test, equal variances assumed, as vector.
```

```
myf.get.p <- function(experiment) {
```

```
  t_value <- t.test(experiment[, 1], experiment[, 2], var.equal = TRUE)[[1]]
```

```
  if (t_value > 0){          # The observed effect is in the wrong direction
```

```
    p <- .999
```

```
  } else {
```

```
    p <- t.test(experiment[, 1], experiment[, 2], var.equal = TRUE)[3]
```

```

    p <- unlist(p)
    p <- p / 2
  }
  return(p)
}

# ---- myf.get.many_p()
# ---- takes 12 data structures: 2 x (DV + 2 moderator subsets) x (with/without outliers)
# ---- returns vector with 12 p-values
# ---- if t-test cannot be executed, p-value of NA returned
myf.get.many_p <- function(list_dvs_moderator_outliers, n) {
  dv1 <- myf.unlist(list_dvs_moderator_outliers, 1, n, 2)
  dv1_trim <- myf.unlist(list_dvs_moderator_outliers, 2, n, 2)
  dv1_mod0 <- myf.unlist(list_dvs_moderator_outliers, 3, n, 2)
  dv1_mod0_trim <- myf.unlist(list_dvs_moderator_outliers, 4, n, 2)
  dv1_mod1 <- myf.unlist(list_dvs_moderator_outliers, 5, n, 2)
  dv1_mod1_trim <- myf.unlist(list_dvs_moderator_outliers, 6, n, 2)
  dv2 <- myf.unlist(list_dvs_moderator_outliers, 7, n, 2)
  dv2_trim <- myf.unlist(list_dvs_moderator_outliers, 8, n, 2)
  dv2_mod0 <- myf.unlist(list_dvs_moderator_outliers, 9, n, 2)
  dv2_mod0_trim <- myf.unlist(list_dvs_moderator_outliers, 10, n, 2)

```

```
dv2_mod1 <- myf.unlist(list_dvs_moderator_outliers, 11, n, 2)
dv2_mod1_trim <- myf.unlist(list_dvs_moderator_outliers, 12, n, 2)
p <- rep_len(NA, 12)
for (i in 1:12) {
  experiment <- switch(i,
    dv1,
    dv1_trim,
    dv1_mod0,
    dv1_mod0_trim,
    dv1_mod1,
    dv1_mod1_trim,
    dv2,
    dv2_trim,
    dv2_mod0,
    dv2_mod0_trim,
    dv2_mod1,
    dv2_mod1_trim)
  var_ctrlg <- var(experiment[ , 1], na.rm = TRUE)
  var_expg <- var(experiment[ , 2], na.rm = TRUE)
  if (any(is.na(var_ctrlg), is.na(var_expg)) == T) {
    p[i] <- NA
  }
}
```

```

    } else {
      p[i] <- myf.get.p(experiment)
    }
  }
  return(p)
}

```

```

#####
#  fixed parameters  #
#####

```

```

# ---- nma = number of meta-analyses run for each combination of conditions
# ---- z_crit is criterion for outlier selection
# ---- n_vector contains observed sample sizes (per group) from 150 meta-analyses
nma <- 1000
z_crit <- 2
obsN <- read.xlsx("obsN.xlsx", sheetIndex = 1)
n_vector <- obsN[, 3]

```

```

#####
#  run stuff  #

```

#####

# ---- four parameters are systematically varied (fully crossed) in simulations:

# ---- 1) effect\_size is in Cohen's d units

# ---- 2) tau reflects the SD for the true effect size for studies within the same meta-analysis

# ---- 3) k reflects the number of studies in the meta-analysis

# ---- 4) pb reflects the strength of publication bias (proportion of n.s. results that remain unpublished)

# ---- qrp\_type is random variable that determines individual study's QRP level

# ---- QRP levels are none, moderate, aggressive, following CARTER, SCHÖNBRODT, HILGARD, GERVAIS, submitted

# ---- r reflects correlation between dv1 and dv2 in population, cf. myf.alternative.dv()

# ---- absolute\_max\_n relevant for optional stopping, cf. myf.aggr.qrps()

```
result_ma <- matrix( , nrow = 3 * nma, ncol = 6)
```

```
new_result <- matrix( , nrow = 1, ncol = 21)
```

```
all_results <- matrix(0 , nrow = 1, ncol = 21)
```

```
for (loop_es in 1:4) {
```

```
  effect_size <- switch(loop_es, 0, 0.2, 0.5, 0.8)
```

```
  for (loop_tau in 1:3) {
```

```
    tau <- switch(loop_tau, 0, 0.2, 0.4)
```

```
    for (loop_k in 1:4) {
```

```
      k <- switch(loop_k, 10, 30, 60, 100)
```



```

d <- c(1:k)
n <- c(1:k)
vi <- c(1:k)
for (loop_pb in 1:3) {
  pb <- switch(loop_pb, 0, 0.4, 0.8)
  cat(effect_size, tau, k, pb, "\n")
  for (loop_nma in 1:nma) {
    i <- 1
    while (i <= k) {
      qrp_type <- runif(1, min=0, max=1)
      if (qrp_type < 0.1) {          # ---- this study is conducted with no QRPs
        r <- 1
        absolute_max_n <- 1
        results <- myf.moderate.qrps(absolute_max_n, effect_size, tau)
      } else if (qrp_type < 0.5) {  # ---- this study is conducted with moderate QRPs
        r <- 0.8
        absolute_max_n <- 200
        results <- myf.moderate.qrps(absolute_max_n, effect_size, tau)
      } else {                    # ---- this study is conducted with aggressive QRPs
        r <- 0.8
        absolute_max_n <- 200
      }
    }
  }
}

```

```

    results <- myf.aggr.qrps(absolute_max_n, effect_size, tau)
  }
  d[i] <- unlist(results[1])
  n[i] <- unlist(results[2])
  vi[i] <- 4*(1+(d[i]^2)/8)/n[i]
  p <- unlist(results[3])
  if (p <= 0.05 && abs(d[i]) < 5) { # ---- d > 5 is ignored as rare, unrealistic; extreme ds distort RMSE
    i <- i + 1
  } else {
    if (p < .999 && runif(1, min=0, max=1) > pb && abs(d[i]) < 5) { # ---- effect in the right direction and study not suppressed by publication bias
      i <- i + 1
    }
  }
} # ---- all studies for meta-analysis completed

  for (loop_MA_model in 1:3) {
# ---- run 3 meta-analysis models (DerSimonian-Laird, Hunter-Schmidt, Paule-Mandel)
  MA_model <- switch(loop_MA_model, "DL", "HS", "PM")
  result_ma[loop_nma + ((loop_MA_model - 1) * nma), ] <- myf.ma(d, n, MA_model)
}
} # ---- all meta-analyses for specific parameter combination completed

for (i in 1:3){ # ---- for each of the 3 meta-analysis models ...

```

```

MA_model <- switch(i, "DL", "HS", "PM")
for (loop_result in 1:21) {           # ---- ... determine, for each condition, the average across nma meta-analyses

new_result[loop_result] <- switch(loop_result,
    "high QRPs one-tailed",
    MA_model,
    effect_size,
    tau,
    k,
    pb,
    mean(result_ma[((i-1) * nma + 1):(nma * i), 1]),    # mean d
    mean(result_ma[((i-1) * nma + 1):(nma * i), 2]),    # mean tau
    median(result_ma[((i-1) * nma + 1):(nma * i), 2]),  # median tau
    rmse(result_ma[((i-1) * nma + 1):(nma * i), 2], tau), # rmse tau
    sd(result_ma[((i-1) * nma + 1):(nma * i), 2]),      # SD tau
    skewness(result_ma[((i-1) * nma + 1):(nma * i), 2]), # skewness tau
    kurtosis(result_ma[((i-1) * nma + 1):(nma * i), 2]), # kurtosis tau
    sum(result_ma[((i-1) * nma + 1):(nma * i), 3]) / nma * 100, # capture % of CI for tau2
    mean(result_ma[((i-1) * nma + 1):(nma * i), 4]),    # mean FI
    mean(result_ma[((i-1) * nma + 1):(nma * i), 5]),    # mean I2
    median(result_ma[((i-1) * nma + 1):(nma * i), 5]),  # median I2

```

```

      sd(result_ma[((i-1) * nma + 1):(nma * i), 5]),      # SD I2
      skewness(result_ma[((i-1) * nma + 1):(nma * i), 5]), # skewness I2
      kurtosis(result_ma[((i-1) * nma + 1):(nma * i), 5]), # kurtosis I2
      mean(result_ma[((i-1) * nma + 1):(nma * i), 6]))    # mean n per meta-analysis (across groups)
    }
    all_results <- rbind(all_results, new_result)
  }
} # ---- all levels for pb completed
} # ---- all levels for k completed
} # ---- all levels for tau completed
} # ---- all levels for effect size completed

colnames(all_results) <- c("simulation", "MA model", "ES", "tau", "k", "PB", "d", "mean_tau", "median_tau", "RMSE_tau", "SD_tau", "skewness_tau",
"skurtosis_tau", "capture % CI tau2", "FI", "mean_I2", "median_I2", "SD_I2", "skewness_I2", "kurtosis_I2", "mean_n")
write.xlsx(all_results, file = "new highQRPs oneTailed2.xlsx", sheetName = "high QRPs oneTailed", row.names = FALSE)

```

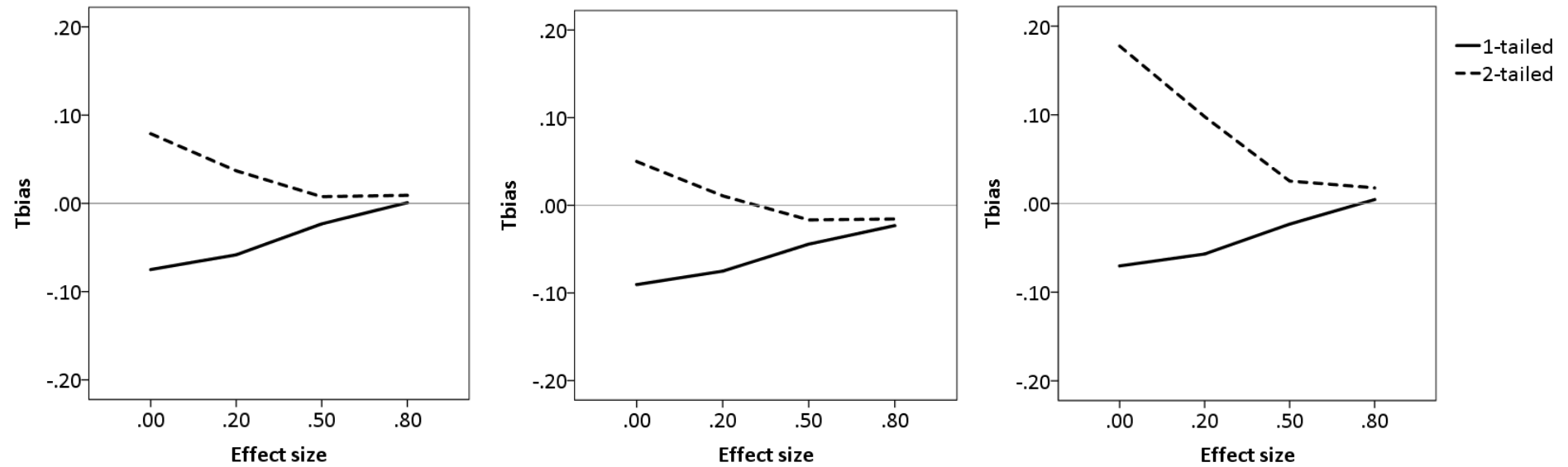
Study 1b: Results based on Der Simonian Laird, Hunter-Schmidt and Paule-Mandel estimators

Effects ( $\eta^2$ ) of simulation parameters on  $T_{\text{bias}}$ .

	$\eta^2$		
	DL	HS	PM
True heterogeneity ( $\tau$ )	0.415	0.499	0.216
One versus two tailed	0.213	0.143	0.275
QRP condition (No, medium, high)	0.063	0.043	0.105
True effect size ( $\delta$ )	0.007	0.005	0.032
Publication bias (PB)	0.004	0.002	0.005
Number of studies ( $k$ )	0.001	0.019	0.000
1 vs 2-tailed x $\delta$	0.133	0.090	0.173
1 vs 2-tailed x $\tau$	0.052	0.033	0.038
$\delta$ x $\tau$	0.034	0.023	0.016
$\delta$ x PB	0.017	0.010	0.019
QRP x $\tau$	0.011	0.006	0.005
QRP x $\delta$	0.006	0.005	0.020
T x $k$	0.005	0.009	0.002
1 vs 2-tailed x QRP	0.003	0.002	0.006
1 vs 2-tailed x PB	0.002	0.001	0.011
1 vs 2-tailed x $k$	0.001	0.000	0.000
$\delta$ x $k$	0.001	0.001	0.000
T x PB	0.001	0.001	0.000
QRP x $k$	0.000	0.000	0.000
QRP x PB	0.000	0.000	0.000
$k$ x PB	0.000	0.000	0.000

*Note:* Effect size calculations were based on  $N = 2$  per cell, where each  $N$  represents the average of 1,000 simulations. Effects are presented in order of magnitude (within main effects and interaction effects).

Study 1b: Interaction graphs for  $T_{\text{bias}}$  based on DerSimonian-Laird, Hunter-Schmidt and Paule-Mandel estimators



$T_{\text{bias}}$  for 1-tailed and 2-tailed testing, across different levels of effect size. Left-hand panel based on DL estimator; middle panel based on HS estimator, right-hand panel based on PM estimator

## Appendix C

### Study 2: List of PsycINFO categories used

<b>PsycINFO code</b>	<b>Sub-discipline and included topics</b>
2100	<b>General psychology</b> <ul style="list-style-type: none"><li>- History and systems</li></ul>
2200	<b>Psychometrics &amp; statistics &amp; methodology</b> <ul style="list-style-type: none"><li>- Tests &amp; testing</li><li>- Sensory &amp; motor testing</li><li>- Developmental scales &amp; schedules</li><li>- Personality scales &amp; inventories</li><li>- Clinical psychological testing</li><li>- Neuropsychological assessment</li><li>- Health psychology testing</li><li>- Educational measurement</li><li>- Occupational &amp; employment testing</li><li>- Consumer opinion &amp; attitude testing</li><li>- Statistics &amp; mathematics</li><li>- Research methods &amp; experimental design</li></ul>
2300	<b>Human experimental psychology</b> <ul style="list-style-type: none"><li>- Sensory perception</li><li>- Visual perception</li><li>- Auditory &amp; speech perception</li><li>- Motor processes</li><li>- Cognitive processes</li><li>- Learning &amp; memory</li><li>- Attention</li><li>- Motivation &amp; emotion</li><li>- Consciousness states</li><li>- Parapsychology</li></ul>
2400	<b>Animal experimental &amp; comparative psychology</b> <ul style="list-style-type: none"><li>- Learning &amp; motivation</li><li>- Social &amp; instinctive behaviour</li></ul>
2500	<b>Physiological psychology &amp; neuroscience</b> <ul style="list-style-type: none"><li>- Genetics</li></ul>

---

	<ul style="list-style-type: none"> <li>- Neuropsychology</li> <li>- Electrophysiology</li> <li>- Physiological processes</li> <li>- Psychophysiology</li> <li>- Psychopharmacology</li> </ul>
2600	<b>Psychology &amp; the humanities</b> <ul style="list-style-type: none"> <li>- Literature &amp; fine arts</li> <li>- Philosophy</li> </ul>
2700	<b>Communication systems</b> <ul style="list-style-type: none"> <li>- Linguistics &amp; language &amp; speech</li> <li>- Mass media communications</li> </ul>
2800	<b>Developmental psychology</b> <ul style="list-style-type: none"> <li>- Cognitive &amp; perceptual development</li> <li>- Psychosocial &amp; personality development</li> <li>- Gerontology</li> </ul>
2900	<b>Social processes &amp; social issues</b> <ul style="list-style-type: none"> <li>- Social structure &amp; organisation</li> <li>- Religion</li> <li>- Culture &amp; ethnology</li> <li>- Marriage &amp; family</li> <li>- Divorce &amp; remarriage</li> <li>- Childrearing &amp; child care</li> <li>- Political processes &amp; political issues</li> <li>- Sex roles &amp; women's issues</li> <li>- Sexual behaviour &amp; sexual orientation</li> <li>- Drug &amp; alcohol usage (legal)</li> </ul>
3000	<b>Social psychology</b> <ul style="list-style-type: none"> <li>- Group &amp; interpersonal processes</li> <li>- Social perception &amp; cognition</li> </ul>
3100	<b>Personality psychology</b> <ul style="list-style-type: none"> <li>- Personality traits &amp; processes</li> <li>- Personality theory</li> <li>- Psychoanalytic theory</li> </ul>
3200	<b>Psychological &amp; physical disorders</b> <ul style="list-style-type: none"> <li>- Psychological disorders</li> </ul>

---



- 
- Affective disorders
  - Schizophrenia & psychotic states
  - Neuroses & anxiety disorders
  - Personality disorders
  - Behaviour disorders & antisocial behaviour
  - Substance abuse & addiction
  - Criminal behaviour & juvenile delinquency
  - Developmental disorders & autism
  - Learning disorders
  - Mental retardation
  - Eating disorders
  - Speech & language disorders
  - Environmental toxins & health
  - Physical & somatoform & psychogenic disorders
  - Immunological disorders
  - Cancer
  - Cardiovascular disorders
  - Neurological disorders & brain damage
  - Vision & hearing & sensory disorders

3300

**Health & mental health treatment & prevention**

- Psychotherapy & psychotherapeutic counselling
  - Cognitive therapy
  - Behaviour therapy & behaviour modification
  - Group & family therapy
  - Interpersonal & client centred & humanistic therapy
  - Psychoanalytic therapy
  - Clinical psychopharmacology
  - Specialised interventions
  - Clinical hypnosis
  - Self help groups
  - Lay & paraprofessional & pastoral counselling
  - Art & music & movement therapy
  - Health psychology & medicine
  - Behavioural & psychological treatment of physical illnesses
  - Medical treatment of physical illness
-

---

	<ul style="list-style-type: none"> <li>- Promotion &amp; maintenance of health &amp; wellness</li> <li>- Health &amp; mental health services</li> <li>- Outpatient services</li> <li>- Community &amp; social services</li> <li>- Home care &amp; hospice</li> <li>- Nursing homes &amp; residential care</li> <li>- Inpatient &amp; hospital services</li> <li>- Rehabilitation</li> <li>- Drug &amp; alcohol rehabilitation</li> <li>- Occupational &amp; vocational rehabilitation</li> <li>- Speech &amp; language therapy</li> <li>- Criminal rehabilitation &amp; penology</li> </ul>
3400	<b>Professional psychological &amp; health personnel issues</b> <ul style="list-style-type: none"> <li>- Professional education &amp; training</li> <li>- Professional personnel attitudes &amp; characteristics</li> <li>- Professional ethics &amp; standards &amp; liability</li> <li>- Impaired professionals</li> </ul>
3500	<b>Educational psychology</b> <ul style="list-style-type: none"> <li>- Educational administration &amp; personnel</li> <li>- Curriculum &amp; programs &amp; teaching methods</li> <li>- Academic learning &amp; achievement</li> <li>- Classroom dynamics &amp; student adjustment &amp; attitudes</li> <li>- Special &amp; remedial education</li> <li>- Gifted &amp; talented</li> <li>- Educational/vocational counselling &amp; student services</li> </ul>
3600	<b>Industrial &amp; organisational psychology</b> <ul style="list-style-type: none"> <li>- Occupational interests &amp; guidance</li> <li>- Personnel management &amp; selection &amp; training</li> <li>- Personnel evaluation &amp; job performance</li> <li>- Management &amp; management training</li> <li>- Personnel attitudes &amp; job satisfaction</li> <li>- Organisational behaviour</li> <li>- Working conditions &amp; industrial safety</li> </ul>
3700	<b>Sport psychology &amp; leisure</b> <ul style="list-style-type: none"> <li>- Sports</li> </ul>

---

---

	- Recreation & leisure
3800	<b>Military psychology</b>
3900	<b>Consumer psychology</b>
	- Consumer attitudes & behaviour
	- Marketing & advertising
4000	<b>Engineering &amp; environmental psychology</b>
	- Human factors engineering
	- Lifespace & institutional design
	- Community & environmental planning
	- Environmental issues & attitudes
	- Transportation
4100	<b>Intelligent systems</b>
	- Artificial intelligence & expert systems
	- Robotics
	- Neural networks
4200	<b>Forensic psychology &amp; legal issues</b>
	- Civil rights & civil law
	- Criminal law & criminal adjudication
	- Mediation & conflict resolution
	- Crime prevention
	- Police & legal personnel

---

Study 2: Results based on Der Simonian Laird, Hunter-Schmidt and Paule-Mandel estimators

Descriptive statistics for meta-analysis quality and heterogeneity ( $T$ ).

For  $d$  and  $T$ ,  $k = 23$  for the large effects sample, and  $k = 43$  for the control sample.

Sample	Quality score	$d$			$T$		
		$DL$	$HS$	$PM$	$DL$	$HS$	$PM$
Large effects	2.43	2.28	2.25	2.34	0.79	0.66	1.15
( $k = 43$ )	(0.98)	(1.55)	(1.53)	(1.56)	(0.43)	(0.36)	(0.91)
Control sample	2.22	0.43	0.42	0.44	0.33	0.28	0.42
( $k = 100$ )	(1.01)	(0.42)	(0.41)	(0.44)	(0.31)	(0.27)	(0.49)

---

<i>All</i>	2.29	0.48	0.41	0.66
	(1.00)	(0.41)	(0.35)	(0.74)

T-tests for differences in heterogeneity ( $T$ ) between groups:

DL:  $t(63) = 4.97, p < .001$

HS:  $t(63) = 4.74, p < .001$

PM:  $t(63) = 4.21, p < .001$

### Regressions for $d$ predicting $T$

Total sample:

DL:  $d$  significantly predicts  $T$ ,  $R = .697, F(1,63) = 59.443, p < .001, T = 0.171 + 0.329d$

HS:  $d$  significantly predicts  $T$ ,  $R = .664, F(1,63) = 49.712, p < .001, T = 0.154 + 0.274d$

PM:  $d$  significantly predicts  $T$ ,  $R = .655, F(1,63) = 47.404, p < .001, T = 0.141 + 0.534d$

Control sample:

DL:  $d$  significantly predicts  $T$ ,  $R = .610, F(1,41) = 24.327, p < .001, T = 0.138 + 0.446d$

HS:  $d$  significantly predicts  $T$ ,  $R = .602, F(1,41) = 23.245, p < .001, T = 0.114 + 0.400d$

PM:  $d$  significantly predicts  $T$ ,  $R = .634, F(1,41) = 27.517, p < .001, T = 0.106 + 0.711d$

Very-large-effect sample:

DL:  $d$  significantly predicts  $T$ ,  $R = .483, F(1,20) = 6.080, p = .023, T = 0.051 + 0.373d$

HS:  $d$  does not significantly predict  $T$ ,  $R = .395, F(1,20) = 3.691, p = .069, T = 0.134 + 0.271d$

PM:  $d$  significantly predicts  $T$ ,  $R = .483, F(1,20) = 6.092, p = .023, T = -0.321 + 0.720d$

Statistics for comparing regression slopes for  $d$  predicting  $T$

Sample	DL		HS		PM	
	$b$	$SE$	$b$	$SE$	$b$	$SE$
All	0.329	0.043	0.274	0.039	0.534	0.078
control	0.446	0.090	0.400	0.083	0.711	0.136
Very large effects	0.373	0.151	0.271	0.141	0.720	0.292

DL: All versus control  $Z = -1.17$ ,  $p = .241$ ; All versus large  $Z = -0.28$ ,  $p = .779$ ; control versus large  $Z = 0.42$ ,  $p = .678$

HS: All versus control  $Z = -1.37$ ,  $p = .169$ ; All versus large  $Z = 0.02$ ,  $p = .984$ ; control versus large  $Z = 0.79$ ,  $p = .430$

PM: All versus control  $Z = -1.13$ ,  $p = .259$ ; All versus large  $Z = -0.62$ ,  $p = .538$ ; control versus large  $Z = -0.03$ ,  $p = .978$

Correlations for exploratory analyses on what drives heterogeneity. k = 22 for large sample (one outlier removed), k = 43 for control sample.

Sample	<i>T and d</i>			<i>T and k</i>			<i>T and quality</i>			<i>d and k</i>		
	<i>DL</i>	<i>HS</i>	<i>PM</i>	<i>DL</i>	<i>HS</i>	<i>PM</i>	<i>DL</i>	<i>HS</i>	<i>PM</i>	<i>DL</i>	<i>HS</i>	<i>PM</i>
Very large effects	<i>r</i> = 0.483, <b><i>p</i> = .023</b>	<i>r</i> = .395, <i>p</i> = .069	<i>r</i> = 0.483, <b><i>p</i> = .023</b>	<i>r</i> = -.088 <i>p</i> = .698	<i>r</i> = .049, <i>p</i> = .828	<i>r</i> = -.106 <i>p</i> = .638	<i>r</i> = -.364, <i>p</i> = .095	<i>r</i> = .043, <i>p</i> = .848	<i>r</i> = -.504, <b><i>p</i> = .017</b>	<i>r</i> = -.134 <i>p</i> = .553	<i>r</i> = -.107, <i>p</i> = .636	<i>r</i> = -.142 <i>p</i> = .528
Control sample	<i>r</i> = .610, <b><i>p</i> &lt; .001</b>	<i>r</i> = .602, <b><i>p</i> &lt; .001</b>	<i>r</i> = .634, <b><i>p</i> &lt; .001</b>	<i>r</i> = -.116, <i>p</i> = .461	<i>r</i> = -.037, <i>p</i> = .815	<i>r</i> = -.159, <i>p</i> = .309	<i>r</i> = -.252, <i>p</i> = .103	<i>r</i> = -.209, <i>p</i> = .179	<i>r</i> = -.234, <i>p</i> = .132	<i>r</i> = -.023, <i>p</i> = .882	<i>r</i> = -.013, <i>p</i> = .935	<i>r</i> = -.034, <i>p</i> = .827
Total sample	<i>r</i> = .697, <b><i>p</i> &lt; .001</b>	<i>r</i> = .664, <b><i>p</i> &lt; .001</b>	<i>r</i> = .655, <b><i>p</i> &lt; .001</b>	<i>r</i> = -.087, <i>p</i> = .492	<i>r</i> = .002, <i>p</i> = .897	<i>r</i> = -.113, <i>p</i> = .369	<i>r</i> = -.264, <b><i>p</i> = .033</b>	<i>r</i> = -.193, <i>p</i> = .124	<i>r</i> = -.336, <b><i>p</i> = .006</b>	<i>r</i> = -.040, <i>p</i> = .753	<i>r</i> = -.029, <i>p</i> = .818	<i>r</i> = -.047, <i>p</i> = .712

## Study 2: Reason for classifying meta-analyses as 'questionable' in importance

Meta-analyses for topics with the largest effect sizes, with reasons for classifying meta-analyses as questionable

Meta-analysis topic and reference	<i>d</i>	<i>T</i>	<i>k</i>	Exp or quasi	Quality	Residual <i>T</i>	<i>FI</i>	Reason for classifying importance as 'questionable'
<b>Patients with psychosis have elevated pro-inflammatory cytokine levels</b> (Upthegrove et al., 2014)	2.21	2.45	5	Quasi	1	1.61	18.4	High <i>T</i> relative to effect size, high residual <i>T</i> , high <i>FI</i> , low quality score
<b>Behavioural techniques as successful treatment for Trichotillomania</b> (Slikboer, Nedeljkovic, Bowe, & Moulding, 2017)	1.85	1.19	4	Exp	3	0.45	6.0	High <i>T</i> relative to effect size, <i>FI</i> shows intervention could be detrimental
<b>Bariatric surgery improves mental-health related quality of life</b> (Magallares & Schomerus, 2015)	9.00	7.54	21	Quasi	2	4.66	11.6	High <i>T</i> relative to effect size, high residual <i>T</i> , high <i>FI</i>

<b>Second language acquisition: Processing Instruction benefits grammar acquisition</b> (Shintani, 2014)	2.60	n/a	42	Exp	1.67	-	-	Low quality score, unclear value of findings
<b>Visual input during tactile stimulation enhances tactile appreciation</b> (Eads, Moseley, & Hillier, 2015)	3.31	2.12	3	Exp	1	0.94	5.9	High <i>T</i> relative to effect size, high residual <i>T</i> , low quality score, effect could reverse ( <i>FI</i> )
<b>Relationship between work engagement and burnout</b> (Maricuțoiu, Sulea, & Iancu, 2017)	1.67	n/a	25	Quasi	1.67	-	-	Unclear nature of relationship which does not add much to knowledge, low quality score
<b>Interventions for internet addiction are effective</b> (Chun, Shim, & Kim, 2017)	1.84	1.04	70	Exp	2	0.30	3.8	High <i>T</i> relative to effect size, intervention could have detrimental effect ( <i>FI</i> )
<b>Efficacy of group contingency interventions with children</b>	3.41	2.45	50	Exp	2	1.25	8.2	High <i>T</i> relative to effect size, high residual <i>T</i> , intervention could have



(Little, Akin-Little, & O'Neill, 2015)									opposite effect ( <i>FI</i> )
<b>Age differences in accuracy of the Social Communication Questionnaire in Autism Spectrum Disorder</b> (Chesnut, Wei, Barnard-Brak, & Richman, 2017)	3.85	n/a	12	Quasi	1.67	-	-		Relatively trivial importance of effect, low quality score
<b>Enhanced consent forms improve participant understanding</b> (Nishimura et al., 2013)	1.73	2.49	11	Exp	2	1.79	24.4		High <i>T</i> relative to effect size, high residual <i>T</i> , high <i>FI</i> suggesting effect can reverse and is not well understood
<b>Aerobic exercise improves executive functioning</b> (Barha, Davis, Falck, Nagamatsu, & Liu-Ambrose, 2017)	2.06	n/a	44	Quasi	0.67	-	-		Unclear findings, very low quality score

## Appendix D

### Study 3: Reference list for height data

- Abbie, A. A. (1967). Skinfold thickness in Australian Aborigines. *Archaeology & Physical Anthropology in Oceania*, 2(3), 207-219.
- Auger, F., Jamison, P., Balslev-Jorgensen, J., Lewin, T., De Pena, J., & Skrobak-Kaczynski, J. (1980). Anthropometry of circumpolar populations. In M. Milan (Ed.), *The Human Biology of Circumpolar Populations* (Vol. 21, pp. 213-255). Cambridge: Cambridge University Press.
- Ayatollahi, S., & Carpenter, R. (1993). Height, weight, BMI and weight-for-height of adults in Southern Iran: How should obesity be defined? *Annals of Human Biology*, 20(1), 13-19.
- Barnicot, N., Bennett, F., Woodburn, J., Pilkington, T., & Antonis, A. (1972). Blood pressure and serum cholesterol in the Hadza of Tanzania. *Human Biology*, 44(1), 87-116.
- Barrett, M. J., & Brown, T. (1971). Increase in average height of Australian Aborigines. *Medical Journal of Australia*, 2(23), 1169-1172.
- Bindon, J. R., & Baker, P. T. (1985). Modernization, migration and obesity among Samoan adults. *Annals of Human Biology*, 12(1), 67-76.
- Birkbeck, J., Lee, M., Myers, G. S., & Alfred, B. M. (1971). Nutritional status of British Columbia Indians: II. Anthropometric measurements, physical and dental examinations at Ahousat and Anaham. *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique*, 403-414.
- Brodsky, D. J. (1906). Zur Topographie des weiblichen Körpers nordostsibirischer Völker. *Archiv für Anthropologie, Jahrgang V, neue reiche*, 1(2), 1-63.
- Cavalli-Sforza, L. (1986). Anthropometric Data. In L. Cavalli-Sforza (Ed.), *African Pygmies* (pp. 81-93). Orlando, FL: Academic Press.
- Chai, C. K. (1967). *Taiwan Aborigines: A genetic study of tribal variations*. Cambridge, MA: Harvard University Press.
- Champness, L., Bradley, M. A., & Walsh, R. (1963). A study of the Tolai in New Britain. *Oceania*, 34(1), 66-75.
- Chang, K. S. F. (1969). *Growth and development of Chinese children & youth in Hong Kong*. Hong Kong: University of Hong Kong.
- Clegg, E. (1989). The growth of Melanesian and Indian children in Fiji. *Annals of Human Biology*, 16(6), 507-528.

- Comas, J. (1971). Anthropometric Studies in Latin American Indian Populations. In F. M. Salzano (Ed.), *The ongoing evolution of Latin American Populations*. Springfield, IL: Charles C Thomas.
- Crognier, E. (1979). Natural selection and physical adaptation to a biotype of tropical Africa. In W. A. Stini (Ed.), *Physiological and morphological adaptation and evolution* (pp. 69-80). Mouton: The Hague.
- Díaz Ungría, A. G. d. (1969). El problema de los pigmeos en América. *Anales de Antropología*, 6(1), 41-78.
- Farabee, W. C. (1924). *The Central Caribs* (Vol. 10). The University Museum: Anthropological Publications.
- Fetter, V., & Hainis, K. (1962). Basic body dimensions of adults of the second Spartakiade. *Acta Universitatis Carolinae Medica.*, 1, 13-31.
- Field, H. (1952). The physical anthropology of the Kurds, Assyrians, Jews, Armenians, Gypsies, and miscellanea. *The anthropology of Iraq* (Vol. XLVI, no 1, pp. 52-63). Harvard University: Papers of the Peabody Museum of American Archaeology and Ethnology.
- Fleischman, M. L. (1980). An unusual distribution for height among males in a Warao Indian village: A possible case of lineal effect. *American Journal of Physical Anthropology*, 53(3), 397-405.
- Fleury-Cuello, E. (1953). Über zwergindianer in Venezuela. *Zeitschrift für Morphologie und Anthropologie* (H. 2), 259-268.
- Friedlaender, J. S. (1987). *The Solomon Islands Project: A long-term study of health, human biology, and culture change*. Oxford: Clarendon Press.
- Frisancho, A. R., & Baker, P. T. (1970). Altitude and growth: A study of the patterns of physical growth of a high altitude Peruvian Quechua population. *American Journal of Physical Anthropology*, 32(2), 279-292.
- Frisancho, A. R., Borkan, G. A., & Klayman, J. E. (1975). Pattern of growth of lowland and highland Peruvian Quechua of similar genetic composition. *Human Biology*, 47(3), 233-243.
- Froment, A., & Hiernaux, J. (1984). Climate-associated anthropometric variation between populations of the Niger bend. *Annals of Human Biology*, 11(3), 189-200.
- Gillin, J. P. (1936). *The Barama River Caribs of British Guiana* (Vol. 14): The Museum.
- Glanville, E., & Geerdink, R. (1970). Skinfold thickness, body measurements and age changes in Trio and Wajana Indians of Surinam. *American Journal of Physical Anthropology*, 32(3), 455-461.
- Gustafsson, A., & Lindenfors, P. (2009). Latitudinal patterns in human stature and sexual stature dimorphism. *Annals of Human Biology*, 36(1), 74-87.

- Harrison, G. A. (1977). *Population structure and human variation* (Vol. 11): Cambridge University Press.
- Hawley, T., & Jansen, A. (1971). Height and weight of Fijians in coastal areas from one year till adulthood. *New Zealand Medical Journal*, 73(469), 346-349.
- Hiernaux, J. (1968). Variabilite du dimorphisme sexuel de la stature en Afrique Subsaharienne et en Europe. In K. Saller (Ed.), *Anthropologie und humangenetik* (pp. 42-50). Stuttgart: Gustav Fisher.
- Hooton, E. A., & Dupertuis, C. W. (1955). *The physical anthropology of Ireland*: The Museum.
- Hyndman, D. C., Ulijaszek, S. J., & Lourie, J. A. (1989). Variability in body physique, ecology, and subsistence in the Fly River region of Papua New Guinea. *American Journal of Physical Anthropology*, 79(1), 89-101.
- Janes, M. (1970). The effect of social class on the physical growth of Nigerian Yoruba children. *Bulletin of the International Epidemiological Association*, 20, 127.
- Khalid, M. (1995). Anthropometric comparison between high-and low-altitude Saudi Arabians. *Annals of Human Biology*, 22(5), 459-465.
- Leonard, W. R., Katzmarzyk, P., Comuzzie, A. G., Crawford, M. H., & Sukernik, R. I. (1994). Growth and nutritional status of the Evenki reindeer herders of Siberia. *American Journal of Human Biology*, 6(3), 339-350.
- Little, M. A., & Johnson Jr, B. R. (1986). Grip strength, muscle fatigue, and body composition in nomadic Turkana pastoralists. *American Journal of Physical Anthropology*, 69(3), 335-344.
- Littlewood, R. (1972). *Physical anthropology of the eastern highlands of New Guinea*: University of Washington Press.
- Malcolm, L. (1969). Determination of the growth curve of the Kukukuku people of New Guinea from dental eruption in children and adult height. *Archaeology & Physical Anthropology in Oceania*, 4(1), 72-78.
- Mueller, W. H., Murillo, F., Palamino, H., Palomino, H., Badzioch, M., Badzioch, M., . . . Schull, W. J. (1980). The Aymara of Western Bolivia: V. Growth and development in an hypoxic environment. *Human Biology*, 52(3), 529-546.
- Neves, W. A., Salzano, F. M., & Da Rocha, F. J. (1985). Principal-components analysis of Brazilian Indian anthropometric data. *American Journal of Physical Anthropology*, 67(1), 13-17.
- Niswander, J., Keiter, F., & Neel, J. (1967). Further studies on the Xavante Indians. II. Some anthropometric, dermatoglyphic, and nonquantitative morphological traits of the Xavantes of Simões Lopes. *American Journal of Human Genetics*, 19(4), 490-501.

- Pretty, G., Henneberg, M., Lambert, K., & Prokopec, M. (1998). Trends in stature in the South Australian Aboriginal Murraylands. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 106(4), 505-514.
- Rosenbaum, S., Skinner, R., Knight, I., & Garrow, J. (1985). A survey of heights and weights of adults in Great Britain, 1980. *Annals of Human Biology*, 12(2), 115-127.
- Rouma, G. (1933). Quitichouas et Aymaras: Étude des populations autochtones des Andes Boliviennes. *Bulletin de la Société Royale Belge d'Anthropologie et de Préhistoire*, 48.
- Schwartz, J., Brumbaugh, R. C., & Chiu, M. (1987). Short stature, growth hormone, insulin-like growth factors, and serum proteins in the Mountain Ok people of Papua New Guinea. *The Journal of Clinical Endocrinology & Metabolism*, 65(5), 901-905.
- Sellen, D. W. (1995). *The socioecology of young child growth among the Datoga pastoralists of Northern Tanzania*. University of California, Davis.
- Shanklin, W. M., & Izzeddin, N. (1937). Anthropology of the near east female. *American Journal of Physical Anthropology*, 22(3), 381-415.
- Shapiro, H. L. (1930). *The physical characters of the Society Islanders* (Vol. 11): Memoirs of the Bernice P Bishop Museum
- Shephard, R. J. (1974). Work physiology and activity patterns of circumpolar Eskimos and Ainu: A synthesis of IBP data. *Human Biology*, 46(2), 263-294.
- Skeller, E. (1954). *Anthropological and ophthalmological studies on the Angmagssalik Eskimos*. Copenhagen: CA Reitzel.
- Slome, C., Gampel, B., Abramson, J., & Scotch, N. (1960). Weight, height and skinfold thickness of Zulu adults in Durban. *South African Medical Journal*, 34(6), 505-509.
- Smirnova, N. (1979). Morphological peculiarities of the body frame in different ethnoterritorial groups. In W. A. Stini (Ed.), *Physiological and morphological adaptation and evolution* (pp. 131-138). Paris: Mouton Publishers
- Spielman, R. S., Da Rocha, F., Weitkamp, L. R., Ward, R. H., Neel, J. V., & Chagnon, N. A. (1972). The genetic structure of a tribal population, the Yanomama Indians VII. Anthropometric differences among Yanomama villages. *American Journal of Physical Anthropology*, 37(3), 345-356.
- Strickland, S., & Ulijaszek, S. (1993). Resting energy expenditure and body composition in rural Sarawaki adults. *American Journal of Human Biology*, 5(3), 341-350.
- Szathmary, E. J., & Holt, N. (1983). Hyperglycemia in Dogrib Indians of the Northwest Territories, Canada: Association with age and a centripetal distribution of body fat. *Human Biology*, 55(2), 493-515.

Valenzuela, C. Y., Rothhammer, F., & Chakraborty, R. (1978). Sex dimorphism in adult stature in four Chilean populations. *Annals of Human Biology*, 5(6), 533-538.

Study 3: Results based on Der Simonian Laird, Hunter-Schmidt and Paule-Mandel estimators

Descriptive statistics for all three heterogeneity estimators (DL = Der Simonian and Laird, HS = Hunter Schmidt, PM = Paule Mandel)

Sex difference (no. countries)	Mean Effect size ( <i>d</i> )			Heterogeneity ( <i>T</i> )			Predicted Heterogeneity ( <i>T</i> )			Residual heterogeneity			Flop Index ( <i>FI</i> )		
	DL	HS	PM	DL	HS	PM	DL	HS	PM	DL	HS	PM	DL	HS	PM
Mate preference: financial (37)	0.738	0.738	0.738	0.328	0.318	0.344	0.175	0.162	0.175	0.153	0.156	0.169	0.01	0.01	0.02
Mate preference: ambition (37)	0.415	0.416	0.419	0.269	0.261	0.236	0.128	0.117	0.128	0.142	0.144	0.108	0.06	0.06	0.04
Mate preference: age difference (37)	2.181	2.178	2.197	0.548	0.530	0.722	0.385	0.364	0.388	0.163	0.166	0.335	0.00	0.00	0.00
Mate preference: looks (37)	- 0.590	-0.589	-0.589	0.195	0.188	0.189	0.153	0.142	0.153	0.042	0.047	0.036	0.00	0.00	0.00
Mate preference: chastity (37)	- 0.323	-0.322	-0.326	0.220	0.213	0.259	0.114	0.104	0.115	0.106	0.108	0.145	0.07	0.06	0.10
Mental rotation: line angle (53)	- 0.510	-0.512	-0.508	0.077	0.064	0.088	0.141	0.131	0.141	- 0.064	-0.066	-0.053	0.00	0.00	0.00

Mental rotation task (52)	- 0.462	-0.464	-0.458	0.042	0.035	0.079	0.134	0.124	0.134	- 0.093	-0.089	-0.054	0.00	0.00	0.00
Mathematics ability at 15: PISA 2015 (69)	- 0.023	-0.023	-0.023	0.048	0.047	0.051	0.070	0.062	0.070	- 0.022	-0.015	-0.019	0.32	0.32	0.33
Mathematics ability 4 <sup>th</sup> grade: TIMSS 2015 (48)	- 0.018	-0.018	-0.018	0.138	0.136	0.144	0.070	0.062	0.070	0.069	0.075	0.075	0.45	0.45	0.45
Mathematics ability 8 <sup>th</sup> grade: TIMSS 2015 (40)	0.027	0.027	0.027	0.104	0.102	0.106	0.070	0.063	0.071	0.033	0.039	0.035	0.40	0.40	0.40
Comparison data															
Height (117)	- 2.009	-2.008	-2.015	0.499	0.482	0.571	0.360	0.340	0.361	0.139	0.141	0.210	0.00	0.00	0.00
2d:4d right hand (42)	0.462	0.462	0.462	0.279	0.267	0.271	0.135	0.124	0.134	0.144	0.143	0.136	0.05	0.04	0.04
Interest in Science: PISA 2006 (57)	- 0.019	-0.019	-0.019	0.112	0.110	0.116	0.070	0.062	0.070	0.042	0.049	0.047	0.43	0.43	0.43
Career in STEM: PISA 2006 (57)	- 0.098	-0.086	-0.086	0.154	0.152	0.160	0.080	0.071	0.080	0.075	0.081	0.080	0.29	0.29	0.30
Gender role attitudes: ISSP 2002 (28)	0.298	0.286	0.286	0.140	0.136	0.141	0.109	0.099	0.109	0.032	0.037	0.032	0.02	0.02	0.02

*Note:* negative Cohen's  $d$  means that men have higher scores on the given sex difference (e.g. men are taller than women, which is reflected in a large, negative Cohen's  $d$ ).



Descriptive statistics for meta-analysis effect size ( $d$ ) and heterogeneity ( $T$ ),  $n = 15$ .

Sample	$d_{HS}$	$d_{DL}$	$d_{PM}$	$T_{HS}$	$T_{DL}$	$T_{PM}$
From $M$ and $SD$	0.543 (0.671)	0.543 (0.671)	0.545 (0.675)	0.203 (0.148)	0.210 (0.153)	0.232 (0.189)
From $N$ and $d$	0.531 (0.640)	0.531 (0.641)	0.533 (0.644)	0.207 (0.158)	0.215 (0.163)	0.239 (0.201)

No significant differences between either  $d$  or  $T$  (for any of the MA models) if calculated from  $M$  and  $SD$  or  $N$  and  $d$ .

Correlations for exploratory analyses on what drives heterogeneity

Sample	$HS$	$DL$	$PM$
$T$ and $d$	$r = .880,$ $p < .001$	$r = 0.887,$ $p < .001$	$r = 0.930,$ $p < .001$
$T$ and $k$	$r = .216,$ $p = .440$	$r = .218,$ $p = .435$	$r = 0.223,$ $p = .424$

#### Regressions for $d$ predicting $T$

HS:  $d$  significantly predicts  $T$ ,  $R = .880$ ,  $F(1,13) = 44.576$ ,  $p < .001$ ,  $T = 0.097 + 0.194d$

DL:  $d$  significantly predicts  $T$ ,  $R = .887$ ,  $F(1,13) = 48.201$ ,  $p < .001$ ,  $T = 0.101 + 0.202d$

PM:  $d$  significantly predicts  $T$ ,  $R = .930$ ,  $F(1,13) = 83.797$ ,  $p < .001$ ,  $T = 0.090 + 0.260d$

### Study 3: Exploratory analysis of use of separate group means ( $N$ for each gender) versus total group size in meta-analysis

In Study 3, we ran meta-analyses to calculate the mean effect size and a measure of heterogeneity ( $T$ ). Using the raw data in this way allows for the variance to be calculated more accurately when group sizes differ (see general methods chapter). In order to establish whether this created any significant differences in mean effect size or heterogeneity, we also conducted the main meta-analyses using Cohen's  $d$  and total group size (the approach that we had to use for Study 1a and Study 2).

We found no significant difference in mean  $d$  or mean  $T$  between analyses conducted based on separate group sizes versus those conducted on total groups. This provides some reassurance that our earlier studies should not be adversely affected by the necessity to conduct analyses based on total group sizes only.

## References

- Abbie, A. A. (1967). Skinfold thickness in Australian aborigines. *Archaeology & Physical Anthropology in Oceania*, 2(3), 207-219.
- Agnihotri, A. K., Jowaheer, A. A., & Soodeen-Lalloo, A. K. (2015). Sexual dimorphism in finger length ratios and sex determination - A study in Indo-Mauritian population. *Journal of Forensic and Legal Medicine*, 35, 45-50. doi:10.1016/j.jflm.2015.07.006
- Almasry, S. M., El Domiaty, M. A., Algaidi, S. A., Elbastawisy, Y. M., & Safwat, M. D. (2011). Index to ring digit ratio in Saudi Arabia at Almadinah Almonawarah province: a direct and indirect measurement study. *Journal of Anatomy*, 218(2), 202-208. doi:10.1111/j.1469-7580.2010.01318.x
- American Psychological Association. (1994). *DSM-IV: Diagnostic and statistical manual of mental disorders* (Vol. 1).
- Anderson, C. A., & Bushman, B. J. (2001). Effects of violent video games on aggressive behavior, aggressive cognition, aggressive affect, physiological arousal, and prosocial behavior: A meta-analytic review of the scientific literature. *Psychological Science*, 12(5), 353-359.
- Andrievskaya, I., & Semenova, M. (2017). Does biological endowment matter for demand for financial services? Evidence from 2D:4D ratio in the Russian household survey. *Personality and Individual differences*, 104, 155-165. doi:10.1016/j.paid.2016.07.019
- Apicella, C. L., Tobolsky, V. A., Marlowe, F. W., & Miller, K. W. (2016). Hadza hunter-gatherer men do not have more masculine digit ratios (2D:4D). *American Journal of Physical Anthropology*, 159(2), 223-232. doi:10.1002/ajpa.22864
- Archer, J. (2019). The reality and evolutionary significance of human psychological sex differences. *Biological Reviews*, 94, 1381-1415.
- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., . . . Nosek, B. A. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27, 108-119.
- Auger, F., Jamison, P., Balslev-Jorgensen, J., Lewin, T., De Pena, J., & Skrobak-Kaczynski, J. (1980). Anthropometry of circumpolar populations. In M. Milan (Ed.), *The Human Biology of Circumpolar Populations* (Vol. 21, pp. 213-255). Cambridge: Cambridge University Press.
- Augusteijn, H. E., van Aert, R., & van Assen, M. A. (2019). The effect of publication bias on the Q test and assessment of heterogeneity. *Psychological Methods*, 24(1), 116-134.

- Ayatollahi, S., & Carpenter, R. (1993). Height, weight, BMI and weight-for-height of adults in southern Iran: How should obesity be defined? *Annals of Human Biology*, 20(1), 13-19.
- Aycinena, D., Baltaduonis, R., & Rentschler, L. (2014). Risk preferences and prenatal exposure to sex hormones for Ladinos. *PloS One*, 9(8). doi:10.1371/journal.pone.0103332
- Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods*, 15(1), 103-133.
- Baker, C. A., Peterson, E., Pulos, S., & Kirkland, R. A. (2014). Eyes and IQ: A meta-analysis of the relationship between intelligence and "Reading the Mind in the Eyes". *Intelligence*, 44, 78-92.
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7(6), 543-554.
- Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, 43(3), 666-678.
- Bandelow, B., Reitt, M., Röver, C., Michaelis, S., Görlich, Y., & Wedekind, D. (2015). Efficacy of treatments for anxiety disorders: A meta-analysis. *International Clinical Psychopharmacology*, 30(4), 183-192.
- Barel, E., & Tzischinsky, O. (2017). The role of sex hormones and of 2D:4D ratio in individual differences in cognitive abilities. *Journal of Cognitive Psychology*, 29(4), 497-507. doi:10.1080/20445911.2017.1279166
- Baribault, B., Donkin, C., Little, D. R., Trueblood, J., Oravecz, Z., van Ravenzwaaij, D., . . . Vandekerckhove, J. (2017). Meta-studies for robust tests of theory. *Proceedings of the National Academy of Sciences*, 115(11), 2607-2612.
- Barnicot, N., Bennett, F., Woodburn, J., Pilkington, T., & Antonis, A. (1972). Blood pressure and serum cholesterol in the Hadza of Tanzania. *Human Biology*, 44(1), 87-116.
- Barrett, M. J., & Brown, T. (1971). Increase in average height of Australian Aborigines. *Medical Journal of Australia*, 2(23), 1169-1172.
- Batz-Barbarich, C., Tay, L., Kuykendall, L., & Cheung, H. K. (2018). A meta-analysis of gender differences in subjective well-being: Estimating effect sizes and associations with gender inequality. *Psychological Science*, 29(9), 1491-1503. doi:10.1177/0956797618774796
- Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25(24), 3186-3191.
- Benton, A. L., Varney, N. R., & deS Hamsher, K. (1978). Visuospatial judgment: A clinical test. *Archives of Neurology*, 35(6), 364-367.

- Berlim, M. T., & Van den Eynde, F. (2014). Repetitive transcranial magnetic stimulation over the dorsolateral prefrontal cortex for treating posttraumatic stress disorder: An exploratory meta-analysis of randomized, double-blind and sham-controlled trials. *The Canadian Journal of Psychiatry*, 59(9), 487-496.
- Bindon, J. R., & Baker, P. T. (1985). Modernization, migration and obesity among Samoan adults. *Annals of Human Biology*, 12(1), 67-76.
- Birkbeck, J., Lee, M., Myers, G. S., & Alfred, B. M. (1971). Nutritional status of British Columbia Indians: II. Anthropometric measurements, physical and dental examinations at Ahousat and Anaham. *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique*, 403-414.
- Bolhuis, J. J., Brown, G. R., Richardson, R. C., & Laland, K. N. (2011). Darwin in mind: New opportunities for evolutionary psychology. *PLoS Biology*, 9(7), e1001109.
- Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester: John Wiley and Sons, Ltd.
- Borenstein, M., Higgins, J., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis:  $I^2$  is not an absolute measure of heterogeneity. *Research Synthesis Methods*, 8(1), 5-18.
- Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, 100(2), 431-449.
- Brodsky, D. J. (1906). Zur topographie des weiblichen körpers Nordostsibirischer völker. *Archiv für Anthropologie, Jahrgang V, neue reiche*, 1(2), 1-63.
- Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, 2(1), 1-38.
- Bushman, B. J., & Anderson, C. A. (2015). Understanding Causality in the Effects of Media Violence. *American Behavioral Scientist*, 59(14), 1807-1821.
- Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences*, 12(1), 1-14.
- Camacho-Hernandez, C. M., Perdigon-Alvarado, E., Quintero-Platt, G., Reyes-Suarez, P., Vera-Delgado, V., Castaneyra-Ruiz, M., . . . Gonzalez-Reimers, E. (2018). Digit ratios among the modern population of the Canary Islands. *European Journal of Anatomy*, 22(2), 145-155.
- Carter, E., Schönbrodt, F., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115-144.
- Cavalli-Sforza, L. (1986). Anthropometric Data. In L. Cavalli-Sforza (Ed.), *African Pygmies* (pp. 81-93). Orlando, FL: Academic Press.

- Chai, C. K. (1967). *Taiwan Aborigines: A genetic study of tribal variations*. Cambridge, MA: Harvard University Press.
- Champness, L., Bradley, M. A., & Walsh, R. (1963). A study of the Tolai in New Britain. *Oceania*, 34(1), 66-75.
- Chang, K. S. F. (1969). *Growth and Development of Chinese Children & Youth in Hong Kong*. Hong Kong: University of Hong Kong.
- Chang, S. H., Stoll, C. R., Song, J., Varela, J. E., Eagon, C. J., & Colditz, G. A. (2014). The effectiveness and risks of bariatric surgery: An updated systematic review and meta-analysis, 2003-2012. *JAMA Surgery*, 149(3), 275-287.
- Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R., Aykutoğlu, B., Bahník, Š., . . . Caprariello, P. (2016). Registered replication report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science*, 11(5), 750-764.
- Chicaiza-Becerra, L. A., & Garcia-Molina, M. (2017). Prenatal testosterone predicts financial risk taking: Evidence from Latin America. *Personality and Individual Differences*, 116, 32-37. doi:10.1016/j.paid.2017.04.021
- Clegg, E. (1989). The growth of Melanesian and Indian children in Fiji. *Annals of Human Biology*, 16(6), 507-528.
- Collaer, M. (2001). Judgment of line angle and position test - 15 line version (JLAP-15). *Unpublished test*.
- Collins, P. Y., Patel, V., Joestl, S. S., March, D., Insel, T. R., Daar, A. S., . . . Fairburn, C. (2011). Grand challenges in global mental health. *Nature*, 475(7354), 27-30.
- Comas, J. (1971). Anthropometric Studies in Latin American Indian Populations. In F. M. Salzano (Ed.), *The ongoing evolution of Latin American Populations*. Springfield, IL: Charles C Thomas.
- Coutrot, A., Silva, R., Manley, E., de Cothi, W., Sami, S., Bohbot, V., . . . Hornberger, M. (2018). Global determinants of navigation ability. *Current Biology*, 28, 1-6.
- Crognier, E. (1979). Natural selection and physical adaptation to a biotype of tropical Africa. In W. A. Stini (Ed.), *Physiological and morphological adaptation and evolution* (pp. 69-80). Mouton: The Hague.
- Darnai, G., Plozer, E., Perlaki, G., Orsi, G., Nagy, S. A., Horvath, R., . . . Clemens, Z. (2016). 2D:4D finger ratio positively correlates with total cerebral cortex in males. *Neuroscience Letters*, 615, 33-36. doi:10.1016/j.neulet.2015.12.056
- Díaz Ungría, A. G. d. (1969). El problema de los pigmeos en América. *Anales de Antropología*, 6(1), 41-78.
- Dieckmann, N. F., Malle, B. F., & Bodner, T. E. (2009). An empirical assessment of meta-analytic practice. *Review of General Psychology*, 13(2), 101-115.

- Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. *American Psychologist*, 54(6), 408-423.
- Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, 6, 621. doi: 10.3389/fpsyg.
- Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Boucher, L. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68-82.
- Eerland, A., Sherrill, A., Magliano, J., Zwaan, R., Arnal, J., Aucoin, P., . . . Carlucci, M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, 11(1), 158-171.
- Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, 136(1), 103-127.
- Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), 591-601.
- Eveleth, P. B. (1975). Differences between ethnic groups in sex dimorphism of adult height. *Annals of Human Biology*, 2(1), 35-39.
- Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS One*, 5(4), e10068.
- Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, 114(14), 3714-3719.
- Fanelli, D., & Ioannidis, J. P. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences*, 110(37), 15031-15036.
- Farabee, W. C. (1924). *The Central Caribs* (Vol. 10). The University Museum: Anthropological Publications.
- Feingold, A. (1994). Gender differences in personality - A meta-analysis. *Psychological Bulletin*, 116(3), 429-456. doi:10.1037/0033-2909.116.3.429
- Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5), 532-538.
- Ferguson, C. J. (2015a). Do Angry Birds make for angry children? A meta-analysis of video game influences on children's and adolescents' aggression, mental health, prosocial behavior, and academic performance. *Perspectives on Psychological Science*, 10(5), 646-666.

- Ferguson, C. J. (2015b). "Everybody knows psychology is not a real science": Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *American Psychologist*, 70(6), 527-542.
- Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, 17(1), 120-128.
- Ferguson, C. J., & Dyck, D. (2012). Paradigm change in aggression research: The time has come to retire the General Aggression Model. *Aggression and Violent Behavior*, 17(3), 220-228.
- Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, 7(6), 555-561.
- Ferguson, C. J., San Miguel, C., & Hartley, R. D. (2009). A multivariate analysis of youth violence and aggression: The influence of family, peers, depression, and media violence. *The Journal of Pediatrics*, 155(6), 904-908.
- Fetter, V., & Hainis, K. (1962). Basic body dimensions of adults of the second Spartakiade. *Acta Universitatis Carolinae Medica*, 1, 13-31.
- Fiedler, K. (2011). Voodoo correlations are everywhere - Not only in neuroscience. *Perspectives on Psychological Science*, 6(2), 163-171.
- Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from  $\alpha$ -error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, 7(6), 661-669.
- Field, H. (1952). The physical anthropology of the Kurds, Assyrians, Jews, Armenians, Gypsies, and miscellanea, *The Anthropology of Iraq* (Vol. XLVI, no 1, pp. 52-63). Harvard University: Papers of the Peabody Museum of American Archaeology and Ethnology.
- Fleischman, M. L. (1980). An unusual distribution for height among males in a Warao Indian village: A possible case of lineal effect. *American Journal of Physical Anthropology*, 53(3), 397-405.
- Fleury-Cuello, E. (1953). Über zwergindianer in Venezuela. *Zeitschrift für Morphologie und Anthropologie*, (H. 2), 259-268.
- Friedlaender, J. S. (1987). *The Solomon Islands Project: A long-term study of health, human biology, and culture change*. Oxford: Clarendon Press.
- Frisancho, A. R., & Baker, P. T. (1970). Altitude and growth: A study of the patterns of physical growth of a high altitude Peruvian Quechua population. *American Journal of Physical Anthropology*, 32(2), 279-292.



- Frisancho, A. R., Borkan, G. A., & Klayman, J. E. (1975). Pattern of growth of lowland and highland Peruvian Quechua of similar genetic composition. *Human Biology*, 47(3), 233-243.
- Froment, A., & Hiernaux, J. (1984). Climate-associated anthropometric variation between populations of the Niger bend. *Annals of Human Biology*, 11(3), 189-200.
- Fusar-Poli, P., Rocchetti, M., Sardella, A., Avila, A., Brandizzi, M., Caverzasi, E., . . . McGuire, P. (2015). Disorder, not just state of risk: Meta-analysis of functioning and quality of life in people at high risk of psychosis. *The British Journal of Psychiatry*, 207(3), 198-206.
- Galeta, P., Bruzek, J., & Laznickova-Galetova, M. (2014). Is sex estimation from handprints in prehistoric cave art reliable? A view from biological and forensic anthropology. *Journal of Archaeological Science*, 45, 141-149. doi:10.1016/j.jas.2014.01.028
- García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research*, 84(1), 74-111.
- Gaulin, S., & Boster, J. (1985). Cross-cultural differences in sexual dimorphism: Is there any variance to be explained? *Ethology and Sociobiology*, 6(4), 219-225.
- Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102, 74-78. doi:10.1016/j.paid.2016.06.069
- Gillin, J. P. (1936). *The Barama River Caribs of British Guiana* (Vol. 14): The Museum.
- Gireesh, A., Das, S., & Viner, R. M. (2018). Impact of health behaviours and deprivation on well-being in a national sample of English young people. *BMJ Paediatrics Open*, 2(1). doi:10.1136/bmjpo-2018-000335
- Glanville, E., & Geerdink, R. (1970). Skinfold thickness, body measurements and age changes in Trio and Wajana Indians of Surinam. *American Journal of Physical Anthropology*, 32(3), 455-461.
- Gleeson, B., & Kushnick, G. (2018). Female status, food security, and stature sexual dimorphism: Testing mate choice as a mechanism in human self-domestication. *American Journal of Physical Anthropology*, 167(3), 458-469.
- Goetzel, R. Z., Long, S. R., Ozminkowski, R. J., Hawkins, K., Wang, S., & Lynch, W. (2004). Health, absence, disability, and presenteeism cost estimates of certain physical and mental health conditions affecting US employers. *Journal of Occupational and Environmental Medicine*, 46(4), 398-412.
- Gonçalves, C., Coelho, T., Machado, S., & Rocha, N. B. (2017). 2D: 4D digit ratio is associated with cognitive decline but not frailty in community-dwelling older adults. *American Journal of Human Biology*, 29(5), e23003.

- Gøtzsche, P. C., Hróbjartsson, A., Marić, K., & Tendal, B. (2007). Data extraction errors in meta-analyses that use standardized mean differences. *JAMA*, 298(4), 430-437.
- Gray, J. P., & Wolfe, L. D. (1980). Height and sexual dimorphism of stature among human societies. *American Journal of Physical Anthropology*, 53(3), 441-456.
- Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*, 7(2), 99-108.
- Grotzinger, J., Jordan, T. H., & Press, F. (2014). *Understanding Earth* (7th ed.): WH Freeman.
- Gustafsson, A., & Lindenfors, P. (2009). Latitudinal patterns in human stature and sexual stature dimorphism. *Annals of Human Biology*, 36(1), 74-87.
- Guttormsen, L. S., Kefalianos, E., & Næss, K.-A. B. (2015). Communication attitudes in children who stutter: A meta-analytic review. *Journal of Fluency Disorders*, 46, 1-14.
- Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A., . . . Bruyneel, S. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546-573.
- Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin*, 136(4), 495-525.
- Harrison, G. A. (1977). *Population structure and human variation* (Vol. 11): Cambridge University Press.
- Hawley, T., & Jansen, A. (1971). Height and weight of Fijians in coastal areas from one year till adulthood. *New Zealand Medical Journal*, 73(469), 346-349.
- Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist*, 42(5), 443-455.
- Hiernaux, J. (1968). Variabilite du dimorphisme sexuel de la stature en Afrique Subsaharienne et en Europe. In K. Saller (Ed.), *Anthropologie und Humangenetik* (pp. 42-50). Stuttgart: Gustav Fisher.
- Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, 327(7414), 557-560.
- Hönekopp, J., & Watson, S. (2010). Meta-analysis of digit ratio 2D: 4D shows greater sex difference in the right hand. *American Journal of Human Biology*, 22(5), 619-630.
- Hönekopp, J., & Watson, S. (2011). Meta-analysis of the relationship between digit-ratio 2D:4D and aggression. *Personality and Individual Differences*, 51(4), 381-386.
- Hooton, E. A., & Dupertuis, C. W. (1955). *The physical anthropology of Ireland*: The Museum.
- Hsu, C. C., Su, B., Kan, N. W., Lai, S. L., Fong, T. H., Chi, C. P., . . . Hsu, M. C. (2015). Elite collegiate tennis athletes have lower 2D:4D ratios than those of nonathlete controls. *Journal of Strength and Conditioning Research*, 29(3), 822-825.

- Huhn, M., Tardy, M., Spineli, L. M., Kissling, W., Förstl, H., Pitschel-Walz, G., . . . Davis, J. M. (2014). Efficacy of pharmacotherapy and psychotherapy for adult psychiatric disorders: A systematic overview of meta-analyses. *JAMA Psychiatry*, 71(6), 706-715.
- Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, 8(1), 3-7.
- Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*: Sage.
- Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, 60(6), 581-592.
- Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, 65, 373-398.
- Hyndman, D. C., Ulijaszek, S. J., & Lourie, J. A. (1989). Variability in body physique, ecology, and subsistence in the Fly River region of Papua New Guinea. *American Journal of Physical Anthropology*, 79(1), 89-101.
- IntHout, J., Ioannidis, J. P., Borm, G. F., & Goeman, J. J. (2015). Small studies are more heterogeneous than large ones: A meta-meta-analysis. *Journal of Clinical Epidemiology*, 68(8), 860-869.
- Ioannidis, J. P. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice*, 14(5), 951-957.
- Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7(6), 645-654.
- Ioannidis, J. P., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal*, 176(8), 1091-1096.
- Irons, W. (1998). Adaptively relevant environments versus the environment of evolutionary adaptedness. *Evolutionary Anthropology*, 6(6), 194-204.
- Jackson, D. (2006). The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine*, 25(17), 2911-2921.
- Jakobsson, U., & Westergren, A. (2005). Statistical methods for assessing agreement for ordinal data. *Scandinavian Journal of Caring Sciences*, 19(4), 427-431.
- Janes, M. (1970). The effect of social class on the physical growth of Nigerian Yoruba children. *Bulletin of the International Epidemiological Association*, 20, 127.
- Jenkins, J. R., Fuchs, L. S., Van Den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95(4), 719-729.
- Jeon, E. H., & Yamashita, J. (2014). L2 Reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64(1), 160-212.

- Jeon, S. W., Yoon, H. K., Han, C., Ko, Y. H., Kim, Y. K., & Won, Y. J. (2016). Second-to-fourth digit length ratio as a measure of harm avoidance. *Personality and Individual Differences*, 97, 30-34.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524-532.
- Johnson, B. T., Scott-Sheldon, L. A., & Carey, M. P. (2010). Meta-synthesis of health behavior change meta-analyses. *American Journal of Public Health*, 100(11), 2193-2198.
- Jonsson, H., Kristensen, M., & Arendt, M. (2015). Intensive cognitive behavioural therapy for obsessive-compulsive disorder: A systematic review and meta-analysis. *Journal of Obsessive-Compulsive and related disorders*, 6, 83-96.
- Joyce, C. W., Mahon, N., Kelly, J. C., Murphy, S., McAllister, M., Chan, J. C., . . . Kelly, J. L. (2014). Hands of a surgeon: Second to fourth digit ratios in the surgical profession. *Personality and Individual differences*, 68, 28-31.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263-291.
- Kang, D., Wu, Y., Hu, D., Hong, Q., Wang, J., & Zhang, X. (2012). Reliability and external validity of AMSTAR in assessing quality of TCM systematic reviews. *Evidence-Based Complementary and Alternative Medicine*, 2012. doi:10.1155/2012/732195
- Kazis, L. E., Anderson, J. J., & Meenan, R. F. (1989). Effect sizes for interpreting changes in health status. *Medical Care*, s178-s189.
- Kedzior, K. K., Gellersen, H. M., Brachetti, A. K., & Berlim, M. T. (2015). Deep transcranial magnetic stimulation (DTMS) in the treatment of major depression: An exploratory systematic review and meta-analysis. *Journal of Affective Disorders*, 187, 73-83.
- Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*. doi:10.1037/met0000209
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196-217.
- Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, 62(6), 593-602.
- Khalid, M. (1995). Anthropometric comparison between high-and low-altitude Saudi Arabians. *Annals of Human Biology*, 22(5), 459-465.

- Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., . . . Brumbaugh, C. C. (2014). Investigating variation in replicability. *Social Psychology*, 45(3), 142-152.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., . . . Bahník, Š. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490.
- Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PloS One*, 9(9), e105825.
- Kyriakidis, I., Papaioannidou, P., Pantelidou, V., Kalles, V., & Gemitzis, K. (2010). Digit ratios and relation to myocardial infarction in Greek men and women. *Gender Medicine*, 7(6), 628-636.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 863.
- Langan, D., Higgins, J., & Simmonds, M. (2015). An empirical comparison of heterogeneity variance estimators in 12,894 meta-analyses. *Research Synthesis Methods*, 6(2), 195-205.
- Langan, D., Higgins, J. P., & Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Research Synthesis Methods*, 8(2), 181-198.
- LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, 15(4), 371-379.
- Leonard, W. R., Katzmarzyk, P., Comuzzie, A. G., Crawford, M. H., & Sukernik, R. I. (1994). Growth and nutritional status of the Evenki reindeer herders of Siberia. *American Journal of Human Biology*, 6(3), 339-350.
- Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs*, 76(3), 286-302.
- Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin and Review*, 21(4), 861-883.
- Lippa, R. A. (2009). Sex differences in sex drive, sociosexuality, and height across 53 nations: Testing evolutionary and social structural theories. *Archives of Sexual Behavior*, 38(5), 631-651.

- Lippa, R. A. (2010). Sex differences in personality traits and gender-related occupational preferences across 53 nations: Testing evolutionary and social-environmental theories. *Archives of Sexual Behavior*, 39(3), 619-636.
- Lippa, R. A., Collaer, M. L., & Peters, M. (2010). Sex differences in mental rotation and line angle judgments are positively associated with gender equality and economic development across 53 nations. *Archives of Sexual Behavior*, 39(4), 990-997.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *American Psychologist*, 48(12), 1181-1209.
- Little, A. C., DeBruine, L. M., & Jones, B. C. (2010). Exposure to visual cues of pathogen contagion changes preferences for masculinity and symmetry in opposite-sex faces. *Proceedings of the Royal Society B: Biological Sciences*, 278(1714), 2032-2039.
- Little, A. C., DeBruine, L. M., & Jones, B. C. (2013). Environment contingent preferences: Exposure to visual cues of direct male-male competition and wealth increase women's preferences for masculinity in male faces. *Evolution and Human Behavior*, 34(3), 193-200.
- Little, M. A., & Johnson Jr, B. R. (1986). Grip strength, muscle fatigue, and body composition in nomadic Turkana pastoralists. *American Journal of Physical Anthropology*, 69(3), 335-344.
- Littlewood, R. (1972). *Physical anthropology of the eastern highlands of New Guinea*: University of Washington Press.
- Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What Is the male-to-female ratio in Autism Spectrum Disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(6), 466-474.
- Lopez, A. D., & Murray, C. C. (1998). The global burden of disease, 1990–2020. *Nature Medicine*, 4(11), 1241-1243.
- Lu, H., Shen, D., Wang, L., Niu, S. B., Bai, C. Y., Ma, Z. B., & Huo, Z. H. (2017). Digit ratio (2D:4D) and handgrip strength are correlated in women (but not in men) in Hui ethnicity. *Early Human Development*, 109, 21-25.
- Lynn, R., & Irwing, P. (2008). Sex differences in mental arithmetic, digit span, and g defined as working memory capacity. *Intelligence*, 36(3), 226-235.
- MacInnes, J. (2005). Work–life balance and the demand for reduction in working hours: Evidence from the British Social Attitudes Survey 2002. *British Journal of Industrial Relations*, 43(2), 273-295.
- Madrian, B. C., & Shea, D. F. (2001). The power of suggestion: Inertia in 401 (k) participation and savings behavior. *The Quarterly Journal of Economics*, 116(4), 1149-1187.

- Magallares, A., & Schomerus, G. (2015). Mental and physical health-related quality of life in obese patients before and after bariatric surgery: A meta-analysis. *Psychology, Health and Medicine, 20*(2), 165-176.
- Malcolm, L. (1969). Determination of the growth curve of the Kukukuku people of New Guinea from dental eruption in children and adult height. *Archaeology & Physical Anthropology in Oceania, 4*(1), 72-78.
- Mann, A., Legewie, J., & DiPrete, T. A. (2015). The role of school performance in narrowing gender gaps in the formation of STEM aspirations: A cross-national study. *Frontiers in Psychology, 6*, 171, doi: 10.3389/fpsyg.2015.00171
- Mann, V. A., Sasanuma, S., Sakuma, N., & Masaki, S. (1990). Sex-differences in cognitive abilities - A cross-cultural perspective. *Neuropsychologia, 28*(10), 1063-1077.
- Marczak, M., Misiak, M., Sorokowska, A., & Sorokowski, P. (2018). No sex difference in digit ratios (2D:4D) in the traditional Yali of Papua and its meaning for the previous hypotheses on the inter-population variability in 2D:4D. *American Journal of Human Biology, 30*(2). e23078 doi:10.1002/ajhb.23078
- Martin, M. O., Mullis, I. V., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015*. Retrieved from: <https://timssandpirls.bc.edu/publications/timss/2015-methods/T15-Methods-and-Procedures-TIMSS-2015.pdf>
- Mccrone, P., Knapp, M., Proudfoot, J., Ryden, C., Cavanagh, K., Shapiro, D. A., . . . Mann, A. (2004). Cost-effectiveness of computerised cognitive-behavioural therapy for anxiety and depression in primary care: Randomised controlled trial. *The British Journal of Psychiatry, 185*(1), 55-62.
- McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science, 9*(6), 612-625.
- McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science, 11*(5), 730-749.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806-834.
- Mitchell, G. (2012). Revisiting truth or triviality the external validity of research in the psychological laboratory. *Perspectives on Psychological Science, 7*(2), 109-117.
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264-269.

- Mueller, W. H., Murillo, F., Palamino, H., Palomino, H., Badzioch, M., Badzioch, M., . . . Schull, W. J. (1980). The Aymara of western Bolivia: V. Growth and development in an hypoxic environment. *Human Biology*, 52(3), 529-546.
- Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., . . . Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1). 0021, doi: 10.1038/s41562-016-0021
- Murphy, K. R. (2017). What inferences can and cannot be made on the basis of meta-analysis? *Human Resource Management Review*, 27, 193-200.
- Neves, W. A., Salzano, F. M., & Da Rocha, F. J. (1985). Principal-components analysis of Brazilian Indian anthropometric data. *American Journal of Physical Anthropology*, 67(1), 13-17.
- Nishimura, A., Carey, J., Erwin, P. J., Tilburt, J. C., Murad, M. H., & McCormick, J. B. (2013). Improving understanding in the research informed consent process: A systematic review of 54 interventions tested in randomized control trials. *BMC Medical Ethics*, 14(1), 28.
- Niswander, J., Keiter, F., & Neel, J. (1967). Further studies on the Xavante Indians. II. Some anthropometric, dermatoglyphic, and nonquantitative morphological traits of the Xavantes of Simões Lopes. *American Journal of Human Genetics*, 19(4), 490-501.
- Normann, N., Emmerik, A. A., & Morina, N. (2014). The efficacy of metacognitive therapy for anxiety and depression: A meta-analytic review. *Depression and Anxiety*, 31(5), 402-411.
- Nye, J. V. C., Androuschak, G., Desierto, D., Jones, G., & Yudkevich, M. (2012). 2D:4D asymmetry and gender differences in academic performance. *PloS One*, 7(10). doi:10.1371/journal.pone.0046319
- Office for National Statistics. (2015). Occupational Pension Schemes Survey, UK: 2015. Retrieved from <https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/pensionssavingsandinvestments/bulletins/occupationalpensionschemessurvey/2015>
- Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. H., Oksanen, M. J., & Suggests, M. (2007). The vegan package. *Community Ecology Package*, 10, 631-637.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), 943-951.
- Organisation for Economic Co-operation and Development. (2017). *PISA 2015 Technical Report*. Retrieved from <http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf>



- Oyeyemi, B. F., Iyiola, O. A., Oyeyemi, A. W., Oricha, K. A., Anifowoshe, A. T., & Alamukii, N. A. (2014). Sexual dimorphism in ratio of second and fourth digits and its relationship with metabolic syndrome indices and cardiovascular risk factors. *Journal of Research in Medical Sciences*, 19(3), 234-239.
- Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531-536.
- Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, 36(4), 859-866.
- Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test-different versions and factors that affect performance. *Brain and Cognition*, 28(1), 39-58.
- Pfennig, B. W. (2015). *Principles of Inorganic Chemistry*: John Wiley & Sons.
- PISA 2015 Database. (2015). Retrieved from: <http://www.oecd.org/pisa/data/2015database/>
- Plomin, R. (1990). The role of inheritance in behavior. *Science*, 248(4952), 183-188.
- Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.
- Pretty, G., Henneberg, M., Lambert, K., & Prokopec, M. (1998). Trends in stature in the South Australian Aboriginal Murraylands. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 106(4), 505-514.
- Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7(4), 331-363.
- Rivas, M. P., Moreira, L. M. A., Santo, L. D. E., Marques, A., El-Hani, C. N., & Toralles, M. B. P. (2014). New studies of second and fourth digit ratio as a morphogenetic trait in subjects with congenital adrenal hyperplasia. *American Journal of Human Biology*, 26(4), 559-561.
- Rosa-Alcázar, A. I., Sánchez-Meca, J., Rosa-Alcázar, Á., Iniesta-Sepúlveda, M., Olivares-Rodríguez, J., & Parada-Navas, J. L. (2015). Psychological treatment of obsessive-compulsive disorder in children and adolescents: A meta-analysis. *The Spanish Journal of Psychology*, 18. e20, 1–22.
- Rosenbaum, S., Skinner, R., Knight, I., & Garrow, J. (1985). A survey of heights and weights of adults in Great Britain, 1980. *Annals of Human Biology*, 12(2), 115-127.
- Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. In S. T. Fiske, D. Schacter, & C. Zahn-Waxler (Eds.), *Annual Review of Psychology* (Vol. 52, pp. 59-82). Palo Alto, CA: Annual Reviews.
- Rouma, G. (1933). Quitichouas et Aymaras: Étude des populations autochtones des Andes Boliviennes. *Bulletin de la Société Royale Belge d'Anthropologie et de Préhistoire*, 48.

- Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward?: A different perspective. *Perspectives on Psychological Science*, 4(4), 435-439.
- Rutledge, T., & Loh, C. (2004). Effect sizes and statistical testing in the determination of clinical significance in behavioral medicine research. *Annals of Behavioral Medicine*, 27(2), 138-145.
- Sako, W., Murakami, N., Izumi, Y., & Kaji, R. (2016). The difference of apparent diffusion coefficient in the middle cerebellar peduncle among parkinsonian syndromes: Evidence from a meta-analysis. *Journal of the Neurological Sciences*, 363, 90-94.
- Sako, W., Murakami, N., Izumi, Y., & Kaji, R. (2017). Usefulness of the superior cerebellar peduncle for differential diagnosis of progressive supranuclear palsy: A meta-analysis. *Journal of the Neurological Sciences*, 378, 153-157.
- Sánchez-Meca, J., Rosa-Alcázar, A. I., Iniesta-Sepúlveda, M., & Rosa-Alcázar, Á. (2014). Differential efficacy of cognitive-behavioral therapy and pharmacological treatments for pediatric obsessive-compulsive disorder: A meta-analysis. *Journal of Anxiety Disorders*, 28(1), 31-44.
- Scheib, J. E., Gangestad, S. W., & Thornhill, R. (1999). Facial attractiveness, symmetry and cues of good genes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 266(1431), 1913-1917.
- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17(4), 551-566.
- Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, 62(5), 529-540.
- Schmidt, F. L., & Oh, I. S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, 4(1), 32-37.
- Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2), 90-100.
- Schmitt, D. P. (2015). The evolution of culturally-variable sex differences: Men and women are not always different, but when they are... it appears not to result from patriarchy or sex role socialization. In S. T. & H. R. (Eds.), *The Evolution of Sexuality* (pp. 221-256): Springer.
- Schwartz, J., Brumbaugh, R. C., & Chiu, M. (1987). Short stature, growth hormone, insulin-like growth factors, and serum proteins in the Mountain Ok people of Papua New Guinea. *The Journal of Clinical Endocrinology & Metabolism*, 65(5), 901-905.

- Schwartz, S. H., & Rubel-Lifschitz, T. (2009). Cross-national variation in the size of sex differences in values: Effects of gender equality. *Journal of Personality and Social Psychology, 97*(1), 171-185.
- Sellen, D. W. (1995). *The socioecology of young child growth among the Datoga pastoralists of northern Tanzania*. University of California, Davis.
- Sells, S. B. (1963). An interactionist looks at the environment. *American Psychologist, 18*(11), 696-702.
- Shackelford, T. K., Schmitt, D. P., & Buss, D. M. (2005). Universal dimensions of human mate preferences. *Personality and Individual Differences, 39*(2), 447-458.
- Shanklin, W. M., & Izzeddin, N. (1937). Anthropology of the near east female. *American Journal of Physical Anthropology, 22*(3), 381-415.
- Shapiro, H. L. (1930). *The physical characters of the Society Islanders* (Vol. 11): Memoirs of the Bernice P Bishop Museum
- Sharif, M. O., Janjua-Sharif, F., Ali, H., & Ahmed, F. (2013). Systematic reviews explained: AMSTAR-how to tell the good from the bad and the ugly. *Oral Health Dental Management, 12*(1), 9-16.
- Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., . . . Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology, 7*(1), 10. doi:10.1186/1471-2288-7-10
- Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J., . . . Boers, M. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology, 62*(10), 1013-1020.
- Shephard, R. J. (1974). Work physiology and activity patterns of circumpolar Eskimos and Ainu: A synthesis of IBP data. *Human Biology, 46*(2), 263-294.
- Shintani, N. (2014). The effectiveness of processing instruction and production-based instruction on L2 grammar acquisition: A meta-analysis. *Applied Linguistics, 36*(3), 306-325.
- Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*(1). doi:10.1146/annurev-psych-122216-011845
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.
- Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76-80.

- Skeller, E. (1954). *Anthropological and ophthalmological studies on the Angmagssalik Eskimos*. Copenhagen: CA Reitzel.
- Skinner, B. F. (1948). 'Superstition' in the pigeon. *Journal of Experimental Psychology*, 38(2), 168-172.
- Slome, C., Gampel, B., Abramson, J., & Scotch, N. (1960). Weight, height and skinfold thickness of Zulu adults in Durban. *South African Medical Journal*, 34(6), 505-509.
- Smirnova, N. (1979). Morphological peculiarities of the body frame in different ethnoterritorial groups. In W. A. Stini (Ed.), *Physiological and Morphological Adaptation and Evolution* (pp. 131-138). Paris: Mouton Publishers
- Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32(9), 752-760.
- Sorokowski, P., Sorokowska, A., Danel, D., Mberira, M. L., & Pokrywka, L. (2012). The second to fourth digit ratio and age at first marriage in semi-nomadic people from Namibia. *Archives of Sexual Behavior*, 41(3), 703-710.
- Spielman, R. S., Da Rocha, F., Weitkamp, L. R., Ward, R. H., Neel, J. V., & Chagnon, N. A. (1972). The genetic structure of a tribal population, the Yanomama Indians VII. Anthropometric differences among Yanomama villages. *American Journal of Physical Anthropology*, 37(3), 345-356.
- Stanley, T., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, 144(12), 1325-1346.
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*, 54(285), 30-34.
- Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64(3), 153-181.
- Stickney, L. T., & Konrad, A. M. (2007). Gender-role attitudes and earnings: A multinational study of married women and men. *Sex Roles*, 57(11-12), 801-811.
- Stoet, G., Bailey, D. H., Moore, A. M., & Geary, D. C. (2016). Countries with higher levels of gender equality show larger national sex differences in mathematics anxiety and relatively lower parental mathematics valuation for girls. *PloS One*, 11(4). e0153857.
- Strickland, S., & Ulijaszek, S. (1993). Resting energy expenditure and body composition in rural Sarawaki adults. *American Journal of Human Biology*, 5(3), 341-350.
- Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, 9(1), 59-71.
- Szathmary, E. J., & Holt, N. (1983). Hyperglycemia in Dogrib Indians of the Northwest Territories, Canada: Association with age and a centripetal distribution of body fat. *Human Biology*, 55(2), 493-515.

- Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, 61(6), 401-409.
- te Nijenhuis, J., Jongeneel-Grimen, B., & Armstrong, E. L. (2015). Are adoption gains on the g factor? A meta-analysis. *Personality and Individual Differences*, 73, 56-60.
- te Nijenhuis, J., Jongeneel-Grimen, B., & Kirkegaard, E. O. (2014). Are Headstart gains on the g factor? A meta-analysis. *Intelligence*, 46, 209-215.
- TIMSS 2015 International Database. (2015). Retrieved from:  
<https://timssandpirls.bc.edu/timss2015/international-database/>
- Tversky, A., & Shafir, E. (1992). Choice under conflict: The dynamics of deferred decision. *Psychological Science*, 3(6), 358-361.
- Uptegrove, R., Manzanares-Teson, N., & Barnes, N. M. (2014). Cytokine function in medication-naïve first episode psychosis: A systematic review and meta-analysis. *Schizophrenia Research*, 155(1), 101-108.
- Valenzuela, C. Y., Rothhammer, F., & Chakraborty, R. (1978). Sex dimorphism in adult stature in four Chilean populations. *Annals of Human Biology*, 5(6), 533-538.
- van Aert, R. C., & Jackson, D. (2018). Multistep estimators of the between - study variance: The relationship with the Paule-Mandel estimator. *Statistics in Medicine*, 37(17), 2616-2629.
- Van Erp, S., Verhagen, J., Grasman, R. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990-2013. *Journal of Open Psychology Data*, 5(1). doi: 10.5334/jopd.33
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., . . . Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1), 55-79.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48.
- Vos, T., Corry, J., Haby, M. M., Carter, R., & Andrews, G. (2005). Cost-effectiveness of cognitive-behavioural therapy and drug interventions for major depression. *Australian and New Zealand Journal of Psychiatry*, 39(8), 683-692.
- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R., . . . Blouin-Hudon, E.-M. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917-928.
- Wakabayashi, A., & Nakazawa, Y. (2010). On relationships between digit ratio (2D:4D) and two fundamental cognitive drives, empathizing and systemizing, in Japanese sample. *Personality and Individual Differences*, 49(8), 928-931.

- Walczyk, J. J. (2000). The interplay between automatic and control processes in reading. *Reading Research Quarterly*, 35(4), 554-566.
- Weisman, O., Pelphrey, K. A., Leckman, J. F., Feldman, R., Lu, Y. F., Chong, A. N., . . . Ebstein, R. P. (2015). The association between 2D:4D ratio and cognitive empathy is contingent on a common polymorphism in the oxytocin receptor gene (OXTR rs53576). *Psychoneuroendocrinology*, 58, 23-32.
- Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (Second ed.). Burlington, MA: Elsevier Academic press.
- Yan, T., Xie, Q., Zheng, Z., Zou, K., & Wang, L. (2017). Different frequency repetitive transcranial magnetic stimulation (rTMS) for posttraumatic stress disorder (PTSD): A systematic review and meta-analysis. *Journal of Psychiatric Research*, 89, 125-135.
- Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist*, 70(1), 10-20.
- Zhao, D., Li, B., Yu, K., & Zheng, L. (2012). Digit ratio (2D: 4D) and handgrip strength in subjects of Han ethnicity: Impact of sex and age. *American Journal of Physical Anthropology*, 149(2), 266-271.
- Zhao, D. P., Yu, K. L., Zhang, X. H., & Zheng, L. B. (2013). Digit ratio (2D:4D) and handgrip strength in Hani ethnicity. *PloS One*, 8(10). doi:10.1371/journal.pone.0077958
- Zwaan, R., Etz, A., Lucas, R. E., & Donnellan, B. (2018). Making replication mainstream. *Behavioural and Brain Sciences*, 41, e120. doi:10.1017/S0140525X17001972